DISENTANGLING DIFFUSION DECODING: SELECTION, PACE, AND NUMERICAL SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion decoders differ in position selection, commitment pace, verification, and numerical semantics, while scoring conventions also affect reported gains. We study these choices under explicit comparison contracts across five models, from three model families, and four benchmarks. In fixed-selector sweeps, doubling commitments per step halves denoiser evaluations and lowers accuracy in all 60 adjacent transitions. Selection gains also depend on pace. Across five controlled comparisons on GSM8K, confidence and two adapted attention-based selectors, ADAS and DAPD, are each regenerated at four and eight commitments under a shared harness and scorer within every checkpoint. On four models, ADAS's advantage over confidence increases 7.7–10.2 percentage points more than DAPD's as pace rises, with 95% confidence intervals excluding zero. These results characterize the adapted policies and do not establish task-general superiority. Replica batching changes 24–89% of raw output sequences across five models. We trace these output differences through termination, scorer normalization, answer extraction, and correctness. Extracted mathematical answers differ on 3.9–28.6% of items, whereas correctness changes on 2.4–7.0%; gains and losses partly offset, and every paired aggregate-accuracy interval contains zero. Supporting experiments show that token persistence does not prove peer reinforcement, while offline prediction and reference agreement do not guarantee effective or efficient decoding. Together, these results show why decoder comparisons must specify selection, pace, numerical execution, scoring, and verification rather than treating an end-to-end gain as a single intervention.

1 INTRODUCTION

Masked diffusion language models generate text by iteratively filling masked positions (Shi et al., 2024; Sahoo et al., 2024; Gong et al., 2022; Li et al., 2022; Bie et al., 2026). Unlike autoregressive models (AR), which generate tokens sequentially (Yang et al., 2025; Touvron et al., 2023; Liu et al., 2024), masked diffusion models predict multiple positions from bidirectional context within a sequence or block and refine the output through repeated denoising steps (Zhong et al., 2026). At each step, the decoder determines which predictions to commit, how many to commit before evaluating the model again, and whether to verify or revise the proposed update. These choices jointly shape computational cost and output quality (Deschenaux & Gulcehre, 2024; Peng et al., 2025; Chang et al., 2026).

When several choices differ between decoders, an end-to-end comparison cannot attribute an observed gain to a single component or separate its individual effect from interactions among components (Zhang et al., 2026; Bertolani et al., 2026). Because each committed update changes the state used by subsequent denoising steps, these components can interact throughout the decoding trajectory. Controlled attribution therefore requires a common decomposition that isolates each intervention and limits the conclusion to what the experiment actually varies.

Our framework. To make these distinctions explicit, we describe a decoder through four components: (i) the *evidence* used to score candidate predictions, (ii) the rule used to *select* positions, (iii) the *pace* at which tokens are committed, and (iv) the *verification or revision* of proposed updates. For each experiment, a *comparison contract* specifies what changes, what remains fixed, which outcome is evaluated, and what conclusion the comparison supports. Figure 1 provides a compact view of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

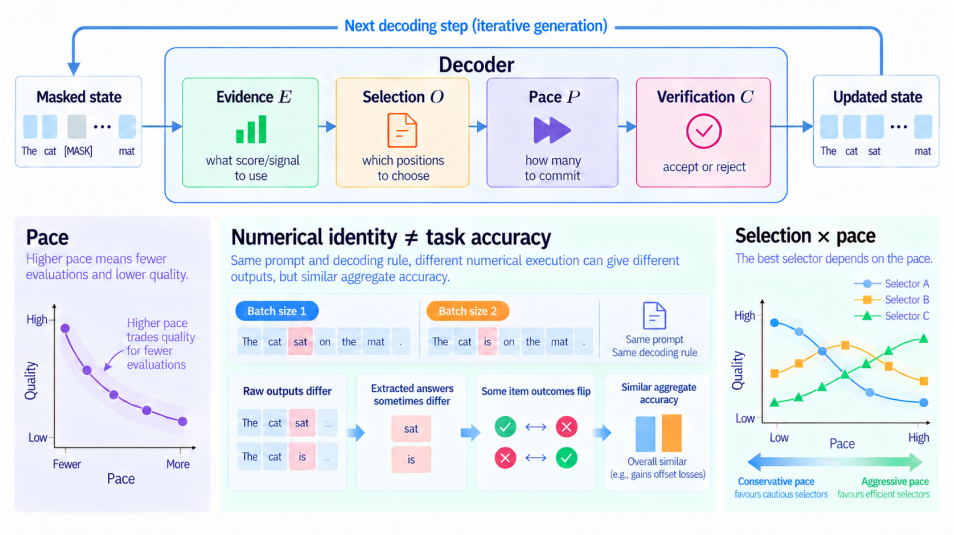

Figure 1: **Decoder factorization and the distinctions.** Each unmasking step uses evidence $\mathcal{E}$ to score candidate predictions, selection $\mathcal{O}$ to choose positions, pace $\mathcal{P}$ to determine how many tokens to commit, and verification $\mathcal{C}$ to accept or revise the update. The lower panels illustrate three distinctions: higher pace trades fewer model evaluations for lower accuracy, numerical execution can change outputs without changing aggregate accuracy, and selector advantage can depend on pace.

this factorization and the central distinctions it enables. We apply the framework across five models from three model families and four benchmarks; the four questions below provide a first view of the resulting study.

Does numerical identity imply task accuracy? *No.* Holding the model, input, and decoding policy fixed, changing execution from batch one to a replica batch of two alters 24–89% of raw generated sequences across five models and four benchmarks, while every paired confidence interval for the aggregate accuracy change includes zero. At the item level, however, 2.4–7.0% of problems switch between correct and incorrect, and extracted mathematical answers change even more frequently. Gains and losses partly offset, so similar aggregate accuracy can conceal differences in generated outputs, extracted answers, and which problems are solved.

Does decoding pace change the benefit of a token-selection rule? *Yes, but with a cost.* Increasing pace reduces computation but can also change the relative value of different selectors. In the fixed-selector sweeps, each doubling of the number of tokens committed per step halves model evaluations but lowers accuracy in every adjacent comparison. Across five controlled GSM8K (Cobbe et al., 2021) comparisons, increasing pace from four to eight tokens per step increases adapted ADAS's (Sahin et al., 2026) advantage over confidence (Wu et al., 2026) more than adapted DAPD's (Kim et al., 2026a) on four of five models. The difference between these changes is 7.7–10.2 percentage points, with 95% confidence intervals excluding zero. Dream-Instruct (Ye et al.) remains unresolved, as does the harmonized LLaDA-1.5/MBPP (Zhu et al., 2026; Austin et al., 2021) comparison, showing that selector rankings must be interpreted at a stated pace and task.

Does persistent token preference imply contextual reinforcement? *Not by itself.* Counterfactual forward passes separate persistence caused by a committed token's own visibility from contextual support provided by its committed peers. Across all five models, committed peers increase the logit margin of proposals that agree with a sequential confidence reference and decrease it for proposals that disagree. Direct visibility also contributes to persistence on the four LLaDA (Zhu et al., 2026; Nie et al., 2026) and Dream models (Ye et al.). These diagnostics therefore show that persistent preference can reflect self-visibility and that peer support is associated with reference agreement, not necessarily task correctness.

Are offline and local diagnostics sufficient to establish deployed policy value? *No.* A learned ranker achieves pooled validation AUROCs of 0.80–0.81 but trails confidence selection by 11–67

accuracy points when used during generation. Conversely, adapted DAPD and margin sometimes receive worse accepted-depth scores under a one-token confidence reference while improving end-task accuracy. Reference fidelity is also distinct from efficiency: on HumanEval (Chen et al., 2021), the tested serial verifier reproduces every reference token sequence across all five models but requires 2.33–3.41 times as many model evaluations. Offline discrimination, local reference agreement, task accuracy, and computational cost therefore require separate evaluation.

Related work. Section 8 positions our contributions against the most closely related work on model families, selection–pace interactions, numerical reproducibility, and representative decoding methods. Appendix B reviews broader decoding methods and benchmark context.

In short our contributions can be summarized as follows:

(i) A factorized framework for decoder evaluation. We describe diffusion decoding through evidence, selection, pace, and verification, and introduce comparison contracts that specify the intervention, controlled conditions, evaluated outcome, and supported conclusion. **(ii) Numerical identity is distinct from task accuracy.** Across five models and four benchmarks, we trace execution-induced differences from raw token sequences through termination, scorer normalization, answer extraction, and correctness, showing that similar aggregate accuracy can conceal different outputs, answers, and item-level outcomes. **(iii) Selection gains depend on decoding pace.** Fixed-selector sweeps and controlled selection–pace comparisons show that adapted selectors respond differently to the same change in the number of tokens committed per step. **(iv) Token persistence does not by itself establish contextual reinforcement.** Counterfactual forward passes separate persistence caused by a token's own visibility from support provided by its committed peers. **(v) Local diagnostics and reference fidelity do not determine deployed policy value.** Learned-ranking, one-action rollout, and verification experiments separate offline discrimination, reference agreement, task performance, and computational efficiency.

2 Decoder factors and comparison contracts

We describe a decoding step through four functional components: evidence $\mathcal{E}$ scores candidate predictions, selection $\mathcal{O}$ chooses masked positions, pace $\mathcal{P}$ determines how many tokens to commit before the next model evaluation, and verification or revision $\mathcal{C}$ accepts or modifies the proposed update. Let x_t denote the partially filled sequence, M_t its masked positions, and $P_t = f_\theta(x_t)$ the token distributions produced by the frozen denoising model. A decoder $D = (\mathcal{E}, \mathcal{O}, \mathcal{P}, \mathcal{C})$ updates the sequence according to: $k_t = \mathcal{P}(x_t, P_t), \qquad x_{t+1} = \mathcal{C}\big(x_t, \mathcal{O}(\mathcal{E}(x_t, P_t), k_t), P_t\big).$

Here, $\mathcal{P}$ sets the update size k_t, $\mathcal{E}$ produces the evidence used by $\mathcal{O}$ to select positions from M_t, and $\mathcal{C}$ produces the next sequence. This factorization describes functional roles rather than a unique implementation: a learned score followed by top-k selection, for example, forms a single policy even when its evidence and selection components are examined separately.

A *comparison contract* specifies the intervention, the components held fixed, the evaluated outcome, and the conclusion supported by the experiment. Fixed-selector comparisons vary pace, whereas fixed-pace comparisons vary the evidence or selection rule. Numerical comparisons vary an execution setting while preserving the model, inputs, and decoding policy. Verification comparisons evaluate both agreement with a specified reference decoder and the computation required to obtain that agreement. The comparison contracts used throughout the study are summarized in Appendix A Table 3; each specifies the intervention, shared evaluation conditions, and supported conclusion.

We evaluate confidence, top-two margin, entropy, temporal stability, attention interaction, and a learned hidden-state score. Selection rules include confidence, left-to-right, seeded random, DAPD (Kim et al., 2026a), ADAS (Sahin et al., 2026), and DAWN (Luo et al., 2026). At fixed pace, structural sets are capped at k or padded with the highest-confidence remaining positions; these are therefore adapted policies, not reproductions of the original decoders.

State-matched diagnostics compare decisions on shared saved states, whereas policy-matched evaluation runs each policy end to end on the states it induces. Accepted depth measures consecutive matches to a one-token confidence reference and favors confidence by construction; a complementary diagnostic substitutes one action and completes with confidence to measure task outcome. Neither replaces full-policy evaluation.

Table 1: **Numerical differences propagate beyond raw outputs.** Replica batch two versus canonical batch one. Output and answer columns report disagreement rates. Gain/loss counts are batch-two-only correct and canonical-only correct; parentheses give the percentage of items whose correctness changes. Net accuracy is batch two minus batch one, and every paired 95% confidence interval includes zero. Extraction failures remain unclassified. For MATH500, answer disagreement denotes a bidirectional non-match under `math_verify`, not proven mathematical inequivalence.

	Output disagreement (%)			**Answer**	**Correctness**	
Task	Raw	Pre-EOS	Scorer	Differs (%)	Gain/loss (flip %)	Δ acc.
Range across cells	**24.0–88.9**	**24.0–71.3**	**14.8–58.4**	**3.9–28.6**	**2.4–7.0 changed**	-2.4 **to** $+0.8$
LLaDA-1.5						
HumanEval	57.9	57.9	27.4	n/a	1/3 (2.4)	-1.2
GSM8K	33.9	33.9	33.8	4.9	13/22 (2.7)	-0.7
MBPP	27.4	27.4	22.4	n/a	8/7 (3.0)	$+0.2$
MATH500	55.4	55.4	55.4	28.6	14/16 (6.0)	-0.4
Dream-7B						
HumanEval	71.3	71.3	25.6	n/a	4/3 (4.3)	$+0.6$
GSM8K	43.1	43.1	43.1	11.2	36/41 (5.8)	-0.4
MBPP	32.8	32.8	16.4	n/a	9/6 (3.0)	$+0.6$
MATH500	69.8	69.8	50.6	25.8	11/15 (5.2)	-0.8
LLaDA-8B-Instruct						
HumanEval	59.8	59.8	28.1	n/a	4/4 (4.9)	$+0.0$
GSM8K	35.6	35.6	35.5	6.1	21/26 (3.6)	-0.4
MBPP	24.0	24.0	18.6	n/a	8/4 (2.4)	$+0.8$
MATH500	50.6	50.6	50.6	25.8	19/16 (7.0)	$+0.6$
Dream-7B-Base						
HumanEval	81.1	64.0	25.0	n/a	1/3 (2.4)	-1.2
GSM8K	88.9	37.9	37.9	8.0	27/29 (4.2)	-0.2
MBPP	34.6	34.6	14.8	n/a	5/12 (3.4)	-1.4
MATH500	58.8	58.8	45.6	24.4	14/15 (5.8)	-0.2
SDAR-8B						
HumanEval	43.9	43.9	18.3	n/a	3/7 (6.1)	-2.4
GSM8K	50.5	50.2	50.0	3.9	18/14 (2.4)	$+0.3$
MBPP	33.8	32.6	29.8	n/a	16/13 (5.8)	$+0.6$
MATH500	64.6	64.6	58.4	23.4	10/17 (5.4)	-1.4

3 Experimental design

We evaluate five pretrained models from three model families: LLaDA-1.5 (Zhu et al., 2026), LLaDA-8B-Instruct (Nie et al., 2026), Dream-v0-Instruct-7B (Ye et al.), Dream-v0-Base-7B, and SDAR-8B-Chat-b32 (Cheng et al., 2026). The benchmarks comprise HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) for code generation, and GSM8K (Cobbe et al., 2021) and MATH500 (Hendrycks et al., 2021) for mathematical reasoning. Unless explicitly varied, generation uses 256-token windows, blocks of 32, greedy token proposals, bfloat16 precision, and batch size one. We score code by test execution, GSM8K by final-answer matching, and MATH500 using `math_verify`. Benchmark loading and reference answers follow the Language Model Evaluation Harness (Zhang et al., 2026), with the task-specific scoring and normalization procedures described below.

The experiments use two evaluation paths. Native fixed-k sweeps use confidence ordering for LLaDA (Nie et al., 2026; Zhu et al., 2026) and SDAR (Cheng et al., 2026) and entropy ordering for Dream (Ye et al.), whereas fixed-pace selection comparisons use a shared project harness with confidence as the reference policy. Because the paths differ in both harness and selection details, we compare absolute accuracies only within the same path.

Accuracy differences use paired item-bootstrap 95% confidence intervals with 10,000 resamples; interaction estimates jointly resample the same items across all relevant arms. Intervals are nominal, without multiplicity adjustment, and those containing zero are treated as unresolved rather than evidence of equivalence. The primary selection–pace experiments regenerate all six selector–pace

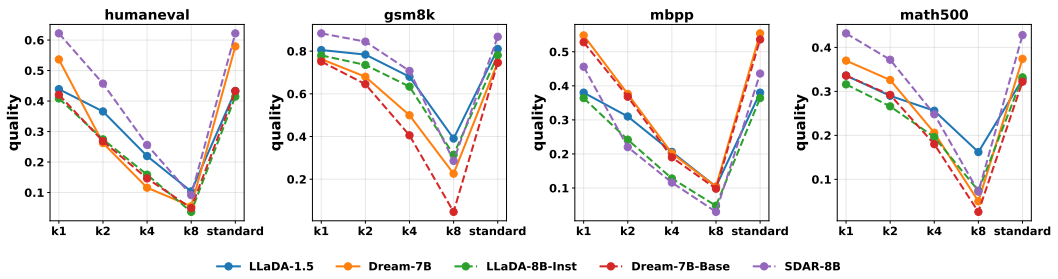

Figure 2: **Accuracy under fixed and adaptive decoding schedules.** Five models and four benchmarks. Fixed paces $k = 1, 2, 4, 8$ require 256, 128, 64, and 32 model evaluations, respectively. Selection is fixed within each fixed-k sweep: confidence for LLaDA and SDAR, and entropy for Dream. Dream's adaptive confidence-threshold comparison changes both selection and pace.

arms under a common harness and scorer within each model and retain completions and raw token sequences. Appendix A documents clustering choices, artifact availability, scoring, termination, numerical settings, and tie handling for every comparison.

4 Numerical execution affects outputs beyond aggregate accuracy

We first test whether deterministic execution choices alter individual decoding decisions. The canonical configuration uses bfloat16, batch size one, a fixed attention backend, and explicit tie conventions. Repeating it is bit-exact, whereas changes to batch shape or composition, attention backend, deterministic kernels, or logit precision alter the selected position in 1.9–4.9% of saved states, primarily near score ties.

We then test whether these local differences propagate through complete generations. The replica-batch experiment processes two copies of the same input together and retains one output while holding the model, input, and decoding policy fixed. Changing from batch size one to two alters 24–89% of raw generated sequences across five models and four benchmarks, although every paired confidence interval for the aggregate accuracy change includes zero. Because this experiment duplicates one request, its measured cost does not represent throughput for batches of distinct requests. Appendix Table 4 reports all twenty model–task comparisons.

Table 1 traces this divergence through raw token arrays, end-token truncation, scorer normalization, answer extraction, and correctness. End-token truncation removes none of the divergence on the three instruction-tuned LLaDA and Dream models, indicating that their differences occur before termination. It removes a large share on Dream-Base—from 88.9% to 37.9% on GSM8K and from 81.1% to 64.0% on HumanEval—and only a small share on SDAR. Task-specific normalization reduces code-output mismatch by 9–64% relative to the pre-termination rate, whereas most mathematical differences survive into the scorer input.

Extracted mathematical answers differ on 3.9–28.6% of items, while correctness changes on 2.4–7.0%. These changes occur in both directions: a *gain* is an item that changes from incorrect under canonical batch-one execution to correct under replica-batch execution, while a *loss* changes from correct to incorrect. For example, LLaDA-1.5/GSM8K has 13 gains and 22 losses. The opposing changes partly offset, so similar aggregate accuracy can conceal differences in generated outputs, extracted answers, and which problems are solved.

Scoring conventions can also change method rankings without changing generations: rescoring the same SDAR/GSM8K DAPD outputs changes its measured difference from confidence from -6.82 to $+3.26$ percentage points (Appendix A.2). Finally, none of the six jointly resampled selector–numerics interaction intervals excludes zero. We therefore detect no change in selector advantage at the available sample sizes, without claiming equality or equivalence across configurations. Appendix C reports the complete intervals, item-level changes, extraction procedures, and interaction estimates.

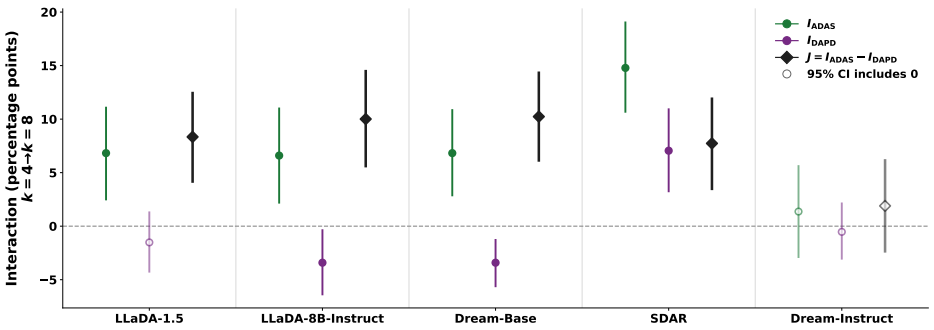

Figure 3: **Selection gains depend on decoding pace.** Points show the jointly resampled change in each selector's advantage over confidence when pace increases from $k = 4$ to $k = 8$ on GSM8K. I_{ADAS} and I_{DAPD} denote the two selector interactions, and $J = I_{\mathrm{ADAS}} - I_{\mathrm{DAPD}}$ compares their responses to pace.

5 Pace and Selection are Distinct but Interacting Factors

Fixed-selector pace effects. We isolate pace by holding each model's selection rule fixed and varying the number of tokens committed per step through $k = 1, 2, 4, 8$. With a fixed 256-token generation window, each doubling of k halves the number of model evaluations. Across five models and four benchmarks, accuracy decreases in all 60 adjacent pace comparisons, as shown in Figure 2, with paired 95% confidence intervals zero in 58 cases. Increasing fixed decoding pace therefore produces a consistent trade-off between accuracy and model evaluations within the evaluated policies.

Relative to $k = 1$, the adaptive confidence-threshold schedule uses only 16–60% as many model evaluations. Its accuracy difference remains unresolved on the four LLaDA and Dream models, including the $n = 5{,}000$ MATH comparison, whereas it is 1.6–2.0 percentage points lower on SDAR's GSM8K and MBPP evaluations. For Dream, however, the fixed-k sweep uses entropy ordering while the adaptive schedule uses confidence; this comparison changes selection and pace together and is therefore interpreted only as a comparison between the two complete policies. Appendix D reports the accuracies, evaluation counts, adjacent-transition intervals, and adaptive-versus-$k = 1$ comparisons in Tables 8–10.

Selection gains depend on decoding pace. At a fixed pace, adapted structural selection can improve accuracy over confidence. At $k = 4$, adapted DAPD improves GSM8K accuracy by 3.3–3.7 percentage points on four models, with paired 95% confidence intervals excluding zero; its effect on Dream-Instruct remains unresolved. Appendix Table 12 reports the underlying accuracies, and Appendix F provides the broader fixed-pace comparisons.

We next test whether selector advantages change with pace. Within each checkpoint, we regenerate confidence, adapted ADAS, and adapted DAPD at $k = 4$ and $k = 8$ on GSM8K using a common harness and scorer. Let $Y_i(o, k) \in \{0, 1\}$ denote correctness on item i under selector o at pace k, and let c denote confidence. We define

$$I_o = \frac{1}{n} \sum_{i=1}^{n} \left([Y_i(o, 8) - Y_i(c, 8)] - [Y_i(o, 4) - Y_i(c, 4)] \right), \tag{1}$$

which measures the change in selector o's advantage over confidence from $k = 4$ to $k = 8$. The contrast $J = I_{\mathrm{ADAS}} - I_{\mathrm{DAPD}}$ tests whether the two selectors respond differently to the same change in pace. Confidence intervals jointly resample the same items across all six selector–pace arms; comparing separate significance decisions at each pace would not estimate this interaction.

Figure 3 summarizes the results. ADAS's advantage over confidence increases with pace on four models, with intervals excluding zero; its positive estimate on Dream-Instruct remains unresolved. DAPD's advantage decreases on LLaDA-8B-Instruct and Dream-Base, increases on SDAR, and remains unresolved on LLaDA-1.5 and Dream-Instruct. The direct contrast J is positive on all five models and excludes zero on four, where it ranges from 7.7 to 10.2 percentage points. Thus, on GSM8K, ADAS responds more positively than DAPD to the same increase in pace. This concerns

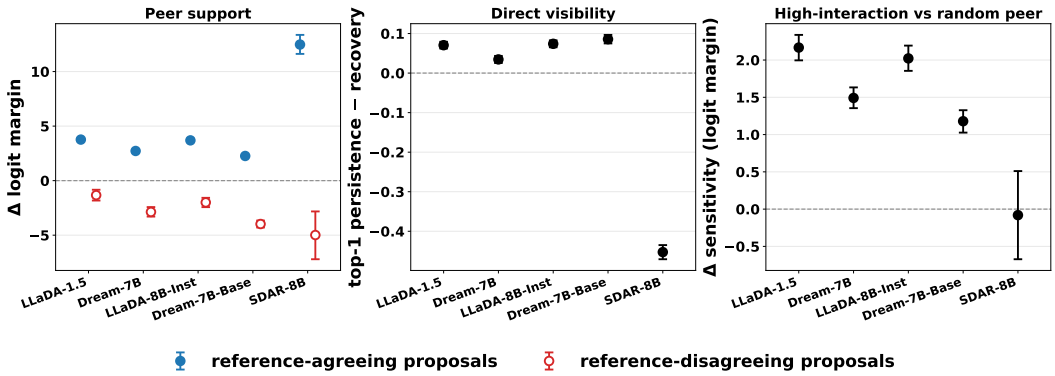

Figure 4: **Counterfactual decomposition of token persistence.** Left: peer-induced change in logit margin, split by agreement with the confidence reference. Middle: direct visibility, measured as the committed state minus leave-one-out recovery. Right: contextual effect of a high-interaction peer relative to a distance-matched random peer. Error bars are state-clustered 95% confidence intervals; the dashed line marks zero. Because SDAR uses a different logit scale, its effects are compared by direction rather than magnitude.

relative advantage: all three policies lose absolute accuracy as pace increases (Appendix Table 15). Exact interaction estimates appear in Appendix Table 11.

The pattern is not established across tasks. In the harmonized LLaDA-1.5/MBPP comparison, both selector interactions and their contrast remain unresolved, with $J = +2.2$ percentage points (95% CI $[-2.6, +7.0]$; Appendix Table 12). Exploratory MBPP comparisons on four additional models likewise resolve no ADAS interaction, although they combine separately dated rather than harmonized runs (Appendix Table 13).

These results characterize the fixed-pace adaptations, not the methods' original decoders. Structural sets larger than k are capped, while shorter sets are padded with confidence-selected positions. At $k = 8$ on LLaDA-1.5/GSM8K, DAPD requires padding on 90% of steps and overlaps confidence's top-k set by 93%, whereas ADAS supplies full-size structural sets. Consequently, J includes the policies' different reliance on confidence padding and does not isolate their native attention mechanisms. Appendix D reports selection sizes and overlap (Table 16) and outcomes across tasks (Table 14).

6 Persistence and Contextual Support

A committed token may remain highly ranked because it is visible at its own position, even when its committed peers provide no contextual support. We separate these effects using counterfactual forward passes on 1,024 saved states per model and candidate groups of size 2, 4, or 8. For each group, we compare the original fully masked state, a state in which the tested token and its peers are committed, and a leave-one-out state in which the peers remain committed but the tested position is remasked. The committed-minus- leave-one-out contrast measures *direct visibility*, while the leave-one-out-minus-original contrast measures *contextual support* from the committed peers.

We also test whether an attention-based interaction signal identifies contextually influential peers by replacing either a high-interaction peer or a distance-matched random peer with its second-ranked token. Proposals are divided by whether they agree with a sequential one-token confidence reference; this measures reference agreement, not ground-truth correctness. Figure 4 summarizes the three tests across all five models, with exact estimates in Appendix Table 17.

Committed peers increase the logit margin of reference-agreeing proposals and decrease it for reference-disagreeing proposals on every model. Peer commitment therefore does not indiscriminately reinforce departures from the reference trajectory, although reference disagreement does not imply an incorrect prediction. Direct visibility also increases persistence on all four LLaDA and Dream models; SDAR differs because visible-position predictions have architecture-specific semantics.

Table 2: **Local diagnostics can disagree with full-policy performance.** Entries count comparisons whose 95% confidence intervals indicate a favorable effect, an unfavorable effect, or remain unresolved. For reference agreement, H/L denotes higher/lower accepted depth than confidence; for outcome evaluations, W/L denotes higher/lower accuracy. The learned ranker achieves pooled validation AUROCs of 0.80 but loses all 20 full-policy comparisons. Reference agreement contains 10 comparisons per non-random rule and three for seeded random; one-action and full-policy evaluations contain five and 20 comparisons per rule, respectively. U denotes an interval containing zero.

Rule	Reference agreement H / L / U	One-action outcome W / L / U	Full policy W / L / U
Adapted DAPD	0/6/4	0/0/5	**6**/1/13
Margin	0/10/0	1/0/4	**3**/1/16
Learned ranker (JUDGE)	0/10/0	0/0/5	0/**20**/0
Seeded random	0/3/0	0/0/5	0/**20**/0

High-interaction peers exert larger contextual effects than matched random peers on those four models, while the SDAR contrast remains unresolved. These diagnostics establish that self-visibility and peer identity affect local predictions, not that they improve end-task accuracy.

Primary intervals resample states. Resampling the 64 source prompts instead widens the intervals but preserves all 20 significance classifications: 19 remain significant, with the same SDAR contrast unresolved (Appendix Table 18). Outcome-stratified analyses likewise associate peer-support direction with reference agreement rather than rollout correctness, but remain descriptive because they condition on completed trajectories (Appendix Table 19).

7 LOCAL DIAGNOSTICS AND VERIFICATION

Local diagnostics are not substitutes for policy evaluation. We use the learned message-passing ranker JUDGE as an instrumented case study, not a proposed decoding method. It aggregates token, distributional, contextual, and hidden-state features, is trained on group-level outcome labels, and is deployed to rank positions within each state (Appendix G). Despite pooled validation AUROCs of 0.80 and 0.81 on LLaDA-1.5 and Dream-Instruct, JUDGE trails confidence by 11–67 accuracy points and loses all 20 policy-matched comparisons. Pooled discrimination across states therefore does not establish effective within-state selection for this ranker, although this does not imply that learned selection policies generally fail.

Reference agreement fails in the opposite direction: adapted DAPD and margin are often penalized for departing from confidence yet produce six and three significant live-policy wins, respectively. One-action rollouts avoid this reference bias but leave 19 of 20 comparisons unresolved. Table 2 summarizes these mismatches. In these experiments, only full-policy evaluation clearly distinguishes useful deviations from policies that consistently reduce accuracy. Complete per-rule and rollout results appear in Appendix F.

Verification requires both fidelity and computational cost. The serial verifier VJUDGE checks confidence-based drafts against a one-token confidence reference and accepts only matching position–token decisions. On full HumanEval, it reproduces every reference token sequence on all five checkpoints, but requires 2.33–3.41 times as many model evaluations. We separately reproduce FREEDAVE (Wu & Zhang, 2025) through the authors' implementation; it decodes faster without a detected accuracy change, but does not reproduce every retained output exactly. These results separate reference fidelity, task accuracy, and computational cost. Appendix G reports the complete verifier procedures and results.

8 RELATED WORK

Diffusion language-model families. Diffusion language models differ in their training objectives and in the sequence structure over which denoising operates. LLaDA (Nie et al., 2026; Zhu et al., 2026; Bie et al., 2026) is trained from scratch with a masked-diffusion objective and repeatedly predicts masked tokens using bidirectional context. Dream (Ye et al.) also performs discrete parallel

denoising, but initializes from an autoregressive model and uses context-adaptive token-level noise scheduling. SDAR (Cheng et al., 2026) instead converts an autoregressive model into a blockwise diffusion model: blocks are generated sequentially, while positions within each block are denoised in parallel. We evaluate two LLaDA models, Dream-Base, Dream-Instruct, and SDAR to test whether decoder-level findings persist across these model constructions.

Decoding methods. Parallel masked-diffusion decoders commonly rank masked positions using token-wise signals such as confidence (Wu et al., 2026), entropy (Ben-Hamu et al., 2026), or probability margin (Zoabi et al., 2026; Kim et al., 2026b), and then commit either a fixed or adaptively chosen number of predictions. Structural selectors additionally model dependencies among candidate positions. DAPD (Kim et al., 2026a) constructs an approximately independent set from attention-based dependencies, whereas ADAS (Sahin et al., 2026) discounts candidates that depend strongly on uncertain selected positions. Our central comparisons use confidence as the reference and adapt DAPD and ADAS to common fixed cardinalities, allowing their selection behavior to be compared at the same pace. Broader decoding approaches are reviewed in Appendix B.

Selection–pace interactions. ADAS preserves its base stopping rule and reports larger benefits under aggressive parallel decoding, while native DAPD couples dependency-aware selection with variable cardinality. We therefore do not claim priority for observing that structural selection and decoding pace can interact. Our contribution is to measure that interaction directly: within each checkpoint, we regenerate confidence, adapted ADAS, and adapted DAPD at $k = 4$ and $k = 8$ under a shared harness and scorer, jointly resample the same items, and compare the selectors through the contrast J. Applying this contract across five models on GSM8K (Cobbe et al., 2021) tests cross-model consistency while retaining an explicitly task-specific scope.

Numerical reproducibility of generation. Shanmugavelu et al. (2024) study floating-point non-associativity and reduction order, while Yuan et al. (2025), Zhang et al. (2025), Gond et al. (2026), and Zhao et al. (2026) examine sensitivity to precision, batching, hardware, tensor parallelism, and kernel execution in language-model inference. Numerical divergence is therefore not unique to diffusion models. We trace how it propagates through iterative diffusion decoding, from position selection and raw tokens through termination, scorer normalization, extracted answers, and correctness, distinguishing output identity, answer agreement, and aggregate accuracy. Additional decoding methods and benchmark context are reviewed in Appendix B.

9 Limitations and conclusion

Limitations. The primary selection–pace result is harmonized across five models but concentrated on GSM8K; the harmonized LLaDA-1.5/MBPP comparison remains unresolved, leaving cross-task generality open. The fixed-pace experiments evaluate capped or confidence-padded adaptations of ADAS and DAPD and therefore characterize these controlled policies rather than the methods' native decoders. The numerical audit spans five models and four benchmarks but tests only a subset of possible hardware, precision, batching, and kernel configurations. The learned ranker and exact verifier are deliberately scoped case studies, while additional SDAR policy arms serve only as sensitivity evidence and are excluded from the primary claims.

Conclusion. This study establishes two main results and two supporting results through explicitly controlled comparison contracts. First, the measured advantage of an adapted token-selection rule depends on commitment pace, and different selectors respond differently to the same pace change. Second, deterministic execution choices can alter generated sequences, extracted answers, and which problems are solved without producing a detected change in aggregate accuracy. The supporting analyses show, first, that persistent token preference does not by itself establish contextual reinforcement and, second, that offline discrimination and reference fidelity do not establish deployed policy value or computational efficiency. Treating selection, pace, scoring, numerical execution, verification semantics, and actual backbone work as part of each reported result makes decoder comparisons easier to attribute and assess.

Reproducibility statement

The appendix specifies generation and scoring conventions, version-dependent corrections, baseline adaptations, and the counting of real backbone evaluations. Aggregate result tables are supplied with the manuscript. Historical model commit revisions and some raw completions were not captured; those gaps are stated rather than inferred. Exact reanalysis requires the separate item-level and prompt-level artifacts. The present manuscript distinguishes completed evidence from measurements whose corrected counterparts have not been verified.

Ethics statement

This study evaluates public language models on established benchmarks without new human-subject data. Agreement with a model's reference output does not certify task correctness or safety. Generated code must be executed in an isolated environment. Numerical output divergence should not be used to imply unmeasured harms or safety guarantees.

Statement of generative AI use

Generative AI tools assisted with language and manuscript editing and with identifying relevant related work. The authors conceived the research, designed and conducted the experiments, performed and verified the analyses, and determined the interpretations and claims. All AI-assisted material, including cited related work, was reviewed and verified by the authors, who take full responsibility for the paper.

References

Sudhanshu Agrawal, Risheek Garrepalli, Raghavv Goel, Christopher Lott, Fatih Porikli, and Mingu Lee. Structuring the future: Diffusion llm speculative decoding via calibrated draft graphs. *arXiv preprint arXiv:2509.18085*, 2025.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Giries Abu Ayoub, Mario Barbara, Lluís Pastor-Pérez, Tanja Bien, Aneesh Barthakur, Alaa Maalouf, and Loay Mualem. Supportive token revealing for fast diffusion language model decoding. *arXiv preprint arXiv:2606.04236*, 2026.

Wenrui Bao, Zhiben Chen, Dan Xu, and Yuzhang Shang. Learning to parallel: Accelerating diffusion large language models via learnable parallel decoding. In *International Conference on Learning Representations*, volume 2026, pp. 83376–83392, 2026.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas S Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *Advances in Neural Information Processing Systems*, 38:55981–56007, 2026.

Thomas Bertolani, Davide Bucciarelli, Leonardo Zini, Marcella Cornia, and Lorenzo Baraldi. Diffusion language models: An experimental analysis. *arXiv preprint arXiv:2606.19475*, 2026.

Tiwei Bie, Maosong Cao, Xiang Cao, Bingsen Chen, Fuyuan Chen, Kun Chen, Lun Du, Daozhuo Feng, Haibo Feng, Mingliang Gong, et al. Llada2. 1: Speeding up text diffusion via token editing. *arXiv preprint arXiv:2602.08676*, 2026.

Tzu-Tao Chang, Benjamin Yuanyang Hong, Kiet Pham, and Shivaram Venkataraman. Dilaserve: High slo attainment serving for diffusion language models. *arXiv preprint arXiv:2606.29094*, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Shuang Cheng, Yihan Bian, Dawei Liu, Yuhua Jiang, Yihao Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, et al. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22058–22075, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Junxia Cui, Haotian Ye, Runchu Tian, Hongcan Guo, Jinya Jiang, Haoru Li, Chaojie Ren, Yiming Huang, Kaijie Zhu, Zhongkai Yu, et al. Simsd: Simple speculative decoding in diffusion language models. *arXiv preprint arXiv:2606.02544*, 2026.

Justin Deschenaux and Caglar Gulcehre. Promises, outlooks and challenges of diffusion language modeling. *arXiv preprint arXiv:2406.11473*, 2024.

Raja Gond, Aditya K Kamath, Ramachandran Ramjee, and Ashish Panwar. Llm-42: Enabling determinism in llm inference with verified speculation. *arXiv preprint arXiv:2601.17768*, 2026.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Michael Hersche, Nicolas Menet, Ronan Tanios, and Abbas Rahimi. Locally coherent parallel decoding in diffusion language models. *arXiv preprint arXiv:2603.20216*, 2026.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating diffusion llms via adaptive parallel decoding. *Advances in neural information processing systems*, 38:52870–52888, 2026.

Metod Jazbec, Theo X Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models. *arXiv preprint arXiv:2512.09106*, 2025.

Bumjun Kim, Dongjae Jeon, Moongyu Jeon, and Albert No. Dapd: Dependency-aware parallel decoding via attention for diffusion llms. *arXiv preprint arXiv:2603.12996*, 2026a.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models. *Advances in Neural Information Processing Systems*, 38:92267–92301, 2026b.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Zirui Wu, Yu Tian, Guorui Zhou, et al. Accelerating diffusion llm inference via local determinism propagation. *arXiv preprint arXiv:2510.07081*, 2025.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Lizhuo Luo, Zhuoran Shi, Jiajun Luo, Zhi Wang, Shen Ren, Wenya Wang, and Tianwei Zhang. Dawn: Dependency-aware fast inference for diffusion llms. *arXiv preprint arXiv:2602.06953*, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Han Peng, Peiyu Liu, Zican Dong, Daixuan Cheng, Junyi Li, Yiru Tang, Shuo Wang, and Wayne Xin Zhao. How efficient are diffusion language models? a critical examination of efficiency evaluation practices. *arXiv preprint arXiv:2510.18480*, 2025.

Liran Ringel, Ameen Ali, and Yaniv Romano. Dependency-guided parallel decoding in discrete diffusion language models. *arXiv preprint arXiv:2604.02560*, 2026.

Yusuf Sahin, Ahmed Rockey Saikia, Volkan Cevher, and Paolo Favaro. Attention-discounted adaptive sampler for masked diffusion language models. *arXiv preprint arXiv:2606.10829*, 2026.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Sanjif Shanmugavelu, Mathieu Taillefumier, Christopher Culver, Oscar Hernandez, Mark Coletti, and Ada Sedova. Impacts of floating-point non-associativity on reproducibility for hpc and deep learning applications. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 170–179. IEEE, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Shengyin Sun, Yiming Li, Renxi Liu, Xinqi Li, Hui-Ling Zhen, Weizhe Lin, Chen Chen, Xianzhi Yu, Mingxuan Yuan, and Chen Ma. Psd: Pushing the pareto frontier of diffusion llms via parallel speculative decoding. *arXiv preprint arXiv:2605.15609*, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. In *International Conference on Learning Representations*, volume 2026, pp. 57027–57051, 2026.

Shutong Wu and Jiawei Zhang. Free draft-and-verification: Toward lossless parallel decoding for diffusion large language models. *arXiv preprint arXiv:2510.00294*, 2025.

Yanzheng Xiang, Lan Wei, Yizhen Yao, Qinglin Zhu, Hanqi Yan, Chen Jin, Philip Alexander Teare, Dandan Zhang, Lin Gui, Amrutha Saseendran, et al. Stop the flip-flop: Context-preserving verification for fast revocable diffusion decoding. *arXiv preprint arXiv:2602.06161*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. *URL https://arxiv. org/abs/2508.15487*.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Give me fp32 or give me death? challenges and solutions for reproducible reasoning. *arXiv e-prints*, pp. arXiv–2506, 2025.

Hengxiang Zhang, Jiaxi Ren, Renchunzi Xie, and Hongxin Wei. Understanding evaluation illusion in diffusion large language models. *arXiv preprint arXiv:2606.29228*, 2026.

Ziyang Zhang, Xinheng Ding, Jiayi Yuan, Rixin Liu, Huizi Mao, Jiarong Xing, and Zirui Liu. Deterministic inference across tensor parallel sizes that eliminates training-inference mismatch. *arXiv preprint arXiv:2511.17826*, 2025.

Shiju Zhao, Jiacheng Yang, Qihang Chen, Junhao Hu, Jiaqi Zheng, Guihai Chen, and Xusheng Chen. Corun: Padding is simple and efficient for deterministic llm inference. *arXiv preprint arXiv:2608.14376*, 2026.

Yangyang Zhong, Yanmei Gu, Zhengqing Zang, Xiaomeng Li, Yuqi Ding, Xibei Jia, Yuting Shen, Zhenzhong Lan, Liwang Zhu, Weiping Liu, et al. Parallelism and generation order in masked diffusion language models: Limits today, potential tomorrow. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7178–7218, 2026.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11425–11460, 2026.

Tamim Zoabi, Ameen Ali, Liran Ringel, and Lior Wolf. Mean-field parallel decoding for discrete diffusion language models. *arXiv preprint arXiv:2606.15805*, 2026.

A ADDITIONAL EXPERIMENTAL DETAILS

Table 3: **Comparison contracts and supported conclusions.** Each contract specifies the intervention, shared evaluation conditions, and scope of the resulting conclusion; empirical coverage varies across contracts.

Comparison	Held fixed / varied	Supported conclusion
Fixed-selector pace	Fix $\mathcal{E}$, $\mathcal{O}$, and $\mathcal{C}$; vary pace $\mathcal{P}$ through k	Effect of pace under the specified selection rule
Fixed-pace selection	Fix k and $\mathcal{C}$; vary $\mathcal{E}$ or $\mathcal{O}$	Effect of the adapted selection rule at the specified pace
Selection–pace interaction	Fix benchmark items, harness, and scorer; vary the selector and k	Whether selector advantage depends on pace and differs between selectors
State-matched diagnostic	Fix the saved state x_t; vary one proposed decision	Local reference agreement or task outcome of one changed decision
Policy-matched evaluation	Fix benchmark items and evaluation settings; run each policy D end to end	Task performance under each policy's induced trajectory
Numerical replay	Fix decoder D and inputs; vary one execution configuration	Sensitivity of generated outputs and task outcomes to numerical execution
Verification audit	Fix the draft and reference policies; measure the work performed by $\mathcal{C}$	Reference fidelity and computational cost of the tested verifier

Table 3 summarizes the comparison contracts used throughout the study. Each contract identifies the intervention, the conditions held fixed, and the conclusion supported by the resulting comparison.

Unless otherwise stated, generation uses a 256-token output window, blocks of 32 positions, greedy token proposals, bfloat16 precision, batch size one, PyTorch 2.7, and CUDA 12.8. Randomness is initialized with Python seed 0 and NumPy/PyTorch seed 1234. Each comparison is additionally specified by its model and tokenizer revision, prompt construction, stopping rules, scorer, tie convention, and evaluation path. Results are compared directly only when the corresponding contract holds these components fixed.

A.1 SCORING AND EVALUATION PATHS

Code scoring. HumanEval uses execution-based pass@1, and MBPP uses its 500-item test split. Code normalization extracts fenced solutions or truncates at a closing fence, detects re-emitted entry-point functions, and evaluates either a complete solution or a continuation as appropriate. This processing materially affects absolute accuracy. On Dream-Instruct HumanEval, accuracy is 15.85% without generated-fence and prose truncation and 53.66% with it. On LLaDA-8B-Instruct HumanEval, interpreting re-emitted function definitions changes the $k = 1$, $k = 4$, and adaptive accuracies from 7.9/1.8/7.3% to 40.9/15.9/41.5%; the corresponding MBPP values change from 29.8/9.6/30.0% to 36.4/12.8/36.4%. The reported results apply the same normalization rule to every retained completion within a comparison.

For the LLaDA-8B-Instruct MBPP $k = 4$ comparisons, the unified scorer yields effects of $+4.4$ points for adapted DAPD and $+6.0$ points for margin; the ADAS effect is invariant to the alternative scoring procedures. Comparisons without retained completion-level outputs remain conditional on their recorded scoring procedure and are not combined with uniformly rescored results.

Evaluation paths. The experiments use two evaluation paths. Native-script fixed-k sweeps use confidence ordering for LLaDA and SDAR and entropy ordering for Dream. Fixed-pace selection comparisons instead use a shared project harness with confidence selection as the reference. Absolute accuracies are therefore compared only within the same evaluation path. For example, Dream-Instruct HumanEval at $k = 4$ obtains 11.6% in the native sweep and 20.1% in the project harness. A direct entropy-versus-confidence comparison identifies the ordering signal as one contributor to this difference but does not provide a complete decomposition of the gap.

Termination handling. Text scoring truncates each generation at its first end-of-sequence or end-of-turn token. This prevents text generated after termination from affecting answer extraction. The

effect is particularly large for Dream-Base/GSM8K at $k = 4$: confidence accuracy is 12.9% when the complete 256-position window is scored and 50.6% after termination-aware truncation, with corresponding reversals in several selector comparisons. On retained $k = 1$ arrays from LLaDA-1.5 and Dream-Instruct, truncation affects only two of 4,966 completions. This comparison characterizes those two sequential policies and does not bound the effect for other models or parallel decoding policies.

GSM8K extraction. GSM8K scoring prioritizes the final boxed expression and requires a digit in fallback numerical matches. This avoids interpreting trailing punctuation or mathematical delimiters as the predicted answer. On retained SDAR generations, applying this extraction rule changes 5,005 scored rows from incorrect to correct and three from correct to incorrect across the evaluated arms; fewer than 20 labels change on each of the other retained checkpoints. Policy comparisons without stored completions remain conditional on their recorded extraction procedure.

MATH500 scoring. MATH500 uses the fixed 500-item subset and evaluates mathematical equivalence with `math_verify`, rather than template exact match. Approximately 5.8% of reference answers cannot be parsed by this scorer and are consequently assigned an incorrect label for every policy. These constant failures contribute zero to paired policy differences under the measured metric, but they do not establish unbiased measurement of mathematical correctness or exclude policy-dependent prediction-format failures. The separate 5,000-item MATH evaluation increases precision for its specified policy comparison; it does not remove the selection–pace confound in Dream's entropy-versus-confidence comparison.

A.2 Tie conventions and SDAR evidence status

SDAR produces exact confidence ties much more frequently than the other evaluated models: 43% of its frozen states contain tied top confidences, compared with 0.1–1.9% on the LLaDA and Dream extension checkpoints. Primary SDAR state-level and verification analyses therefore use an explicit lowest-index tie convention throughout.

To distinguish scoring sensitivity from generation sensitivity, we evaluate three conditions: (A) retained generations under their associated scorer, (B) the same generations under the unified scorer, and (C) generations produced under the explicit tie convention and evaluated by the unified scorer. The A-to-B comparison isolates scoring with generated text fixed, while the B-to-C comparison isolates generation under a fixed scorer. Tables 20 and 21 report condition C for confidence, adapted DAPD, adapted ADAS, and margin at $k = 4$.

The four policies have different relationships to the tie convention:

- *Confidence* uses the explicit lowest-index convention. Across the four tasks, 4 of 2,483 retained completions differ in text between conditions A and C: one on HumanEval and three on GSM8K. The HumanEval item and one GSM8K item retain the same correctness label; the remaining two GSM8K items consist of one gain and one loss. With the scorer fixed, confidence therefore has a net generation effect of 0.0 points on every task.
- *Margin* also uses the explicit convention. Depending on the task, 30–88% of its completions differ between conditions A and C. With scoring held fixed, the corresponding accuracy effect ranges from -0.8 points on MBPP to $+3.4$ points on MATH500. Thus, substantial completion-level divergence can produce a much smaller net accuracy change.
- *DAPD* applies a stable sort to an index-ordered candidate set and therefore resolves ties to the lowest index by construction. Its completions are identical between conditions A and C on all 2,483 items. Its condition-C results consequently reflect scoring differences applied to unchanged generated text.
- *ADAS* delegates selection to a pinned external function whose internal tie convention is not independently verified. Its completions are nevertheless identical between conditions A and C on all 2,483 items. This establishes reproducibility on the evaluated panel, but not compliance with the prescribed lowest-index rule.

The remaining SDAR policy-matched order and evidence arms—left-to-right, seeded random, `judge_local`, and the other evidence signals—are reported only as sensitivity analyses under

their recorded tie behavior and are excluded from the primary structural-selector conclusions. DAWN has no included SDAR arm and is marked unavailable rather than grouped with those sensitivity results.

A.3 Structural-selector implementations

DAPD. The official-code reproduction uses the authors' task-specific thresholds, 256 generated tokens, and a single decoding block. On LLaDA-8B-Instruct, reproduced/published accuracies are 33.5/34.2 on HumanEval, 69.5/71.4 on GSM8K, 35.2/36.0 on MBPP, and 26.2/27.6 on MATH500 under the authors' exact-match metric. The corresponding Dream-Instruct values are 51.8/50.6, 56.7/58.8, 50.2/49.4, and 30.6/30.6. A LLaDA-1.5 implementation obtains 32.9, 70.6, 40.4, and 29.8%, respectively. Because these results use the authors' evaluation path, they are not subtracted from project-harness outcomes to estimate the contribution of an individual decoder component.

The adapted DAPD policy applies the authors' independent-set selection function to attention averaged over the final four layers, rather than using the original multilayer capture. Selected sets larger than k are capped, while shorter sets are padded with the highest-confidence remaining positions. On 64-token development states, mean pre-padding set sizes are 1.04 for LLaDA-1.5 and 1.01 for Dream-Instruct; every $k = 4$ set requires padding and overlaps confidence's top four by approximately 90–92%. In full-length $k = 4$ replica-batch decoding, the mean pre-padding sizes are 4.0 on LLaDA-1.5/GSM8K, 4.2 on Dream-Instruct/GSM8K, and 10.0 on LLaDA-1.5/MBPP, with 82%, 67%, and 40% of steps shorter than k, respectively. The difference reflects where short and full-length trajectories fall within DAPD's threshold schedule. These measurements characterize the fixed-pace adaptation rather than a newly tuned DAPD benchmark result.

ADAS. The official-code Dream-Base reproduction yields a GSM8K advantage of $+6.37$ points at eight commitments per step (95% CI $[4.09, 8.72]$) and a HumanEval advantage of $+8.54$ points at four commitments (95% CI $[0.61, 16.46]$), with identical evaluation counts within each pair. The corresponding published gains are 15.0 and 15.3 points. The reproduction therefore preserves the reported direction but not the full effect magnitude. Because model execution and evaluation settings differ, these results are treated as a reproduction under the specified environment rather than as directly interchangeable with the published values.

The adapted ADAS policy invokes its selection function using attention averaged over the final four layers rather than the original final-layer average and returns exactly k positions. As with adapted DAPD, the reported fixed-pace results characterize this controlled policy rather than the complete native decoder.

SDAR adaptation. SDAR-8B-Chat-b32 uses block-causal attention across 32-position blocks and bidirectional attention within each block. The project adapter left-pads prompts to block alignment, recomputes the sequence without caching, reconstructs attention statistics from query and key projections in the final four layers, and implements normalization through the corresponding PyTorch operations. Its fixed-pace sweep uses the project decoding loop, while its adaptive policy uses a confidence threshold of 0.9. DAPD and ADAS are therefore architectural adaptations on SDAR rather than reproductions of their published implementations. Their selected sets typically contain 7–12 positions on frozen states, making capping more common than confidence-based padding. DAWN is not evaluated on SDAR.

Task-outcome stratification is not informative for SDAR's 64-token development trajectories because these short rollouts rarely solve the underlying tasks and none of the GSM8K trajectories is correct. SDAR also retains its committed argmax at visible positions in approximately 49% of cases, compared with more than 90% for the LLaDA and Dream models. Its visible-token persistence contrast therefore does not measure the same architectural behavior. Masked-position peer-support tests do not rely on this visible-position interpretation and are reported separately.

A.4 Timing and reporting conventions

Selected timing comparisons use three warmed repetitions on fixed 20-item HumanEval subsets, executed sequentially on one A100-40GB GPU. The spread across repetitions is below 2% for all six timed arms, and reported values use the median runtime. These subset measurements assess timing

repeatability under the specified workload; they do not remove possible workload or system variation from other full-benchmark measurements. Replica-batch timing intentionally includes the duplicated computation required to retain one output and should not be interpreted as throughput for batches of distinct requests.

Unless otherwise stated, accuracy is reported in percent and effects in percentage points. Confidence intervals are paired intervals using the sampling procedure specified for each table. A superscript star denotes a nominal interval that excludes zero and does not indicate multiplicity-adjusted evidence. Practical-tie symbols are not used because no formal equivalence margin or equivalence test is specified.

Tables 20 and 21 report the tie-controlled $k = 4$ SDAR results for confidence, adapted DAPD, adapted ADAS, and margin under the arm-specific conditions described above. Other SDAR order and evidence policies are retained only as sensitivity analyses under their recorded conventions, while state-level and verification results use the explicit lowest-index reference. DAWN remains unavailable for SDAR throughout.

B Additional related work

Decoding methods. At each denoising step, a decoder scores the remaining masked positions, selects a subset, and determines how many predictions to commit. Common token-wise signals include confidence, entropy, probability margin, and temporal stability. Dependency-aware methods additionally use attention or token-interaction structure: DAPD (Kim et al., 2026a) selects an approximately independent set of positions, ADAS (Sahin et al., 2026) softly discounts candidates that depend on uncertain selected tokens while retaining the underlying stopping rule, and DAWN (Luo et al., 2026) constructs a dependency graph to avoid jointly committing strongly coupled uncertain positions. In Addition Learn2PD (Bao et al., 2026) trains a lightweight filter to predict token stability and uses it to adapt parallel commitments, while COVER (Xiang et al., 2026) performs context-preserving leave-one-out verification with an adaptive verification budget.

Related methods explore learned position policies (Ringel et al., 2026), structured generation orders, adaptive commitment rules, and draft-and-verify decoding (Israel et al., 2026; Jazbec et al., 2025; Wu & Zhang, 2025; Agrawal et al., 2025; Sun et al., 2026; Hersche et al., 2026; Cui et al., 2026; Zoabi et al., 2026). KLASS (Kim et al., 2026b) uses token-level KL divergence to identify stable predictions for parallel unmasking. LocalLeap (Kong et al., 2025) propagates local determinism from high-confidence anchors and relaxes decoding within bounded neighborhoods, while EB-Sampler (Ben-Hamu et al., 2026) uses an entropy bound to adaptively choose which and how many tokens to unmask. AXON (Ayoub et al., 2026) instead selects confident anchors intended to support the subsequent denoising of uncertain positions. Because many of these methods change selection and pace together, our controlled comparisons hold one component fixed while varying the other and label capped or confidence-padded structural selectors as adapted policies rather than reproductions of their native decoders.

Evaluation benchmarks. Diffusion decoders are commonly evaluated on mathematical reasoning and code generation, where outputs can be scored automatically. GSM8K (Cobbe et al., 2021) contains grade-school mathematical word problems and is scored through extracted final answers, while MATH500 (Hendrycks et al., 2021) contains competition-style problems evaluated through mathematical-answer checking. HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) evaluate program synthesis by executing generated code against unit tests. We use all four benchmarks because they combine different reasoning demands with distinct evaluation pipelines: answer extraction for GSM8K, mathematical-equivalence checking for MATH500, and executable correctness for code. They support precise item-level policy comparisons but do not measure open-ended generation quality.

Table 4: **Full-benchmark replica-batching audit.** Mismatch compares raw token arrays under replica batch two and canonical batch one. Δ accuracy is batch two minus batch one. W/T/L denotes batch-two-only correct, unchanged correctness, and batch-one-only correct. The time ratio includes duplicated computation for one retained output.

Task	n	Mismatch (%)	Δ accuracy [95% CI]	W/T/L	Time ratio
LLaDA-1.5					
HumanEval	164	57.9	$-1.2\,[-3.7, +1.2]$	1/160/3	1.62
GSM8K	1319	33.9	$-0.7\,[-1.6, +0.2]$	13/1284/22	1.86
MBPP	500	27.4	$+0.2\,[-1.4, +1.8]$	8/485/7	1.75
MATH500	500	55.4	$-0.4\,[-2.6, +1.8]$	14/470/16	1.74
Dream-Instruct					
HumanEval	164	71.3	$+0.6\,[-2.4, +3.7]$	4/157/3	1.58
GSM8K	1319	43.1	$-0.4\,[-1.7, +0.9]$	36/1242/41	1.90
MBPP	500	32.8	$+0.6\,[-1.0, +2.2]$	9/485/6	1.83
MATH500	500	69.8	$-0.8\,[-2.8, +1.2]$	11/474/15	1.78
LLaDA-8B-Instruct					
HumanEval	164	59.8	$+0.0\,[-3.0, +3.7]$	4/156/4	1.66
GSM8K	1319	35.6	$-0.4\,[-1.4, +0.6]$	21/1272/26	1.93
MBPP	500	24.0	$+0.8\,[-0.6, +2.2]$	8/488/4	1.83
MATH500	500	50.6	$+0.6\,[-1.8, +2.8]$	19/465/16	1.81
Dream-Base					
HumanEval	164	81.1	$-1.2\,[-3.7, +1.2]$	1/160/3	1.56
GSM8K	1319	88.9	$-0.2\,[-1.3, +1.0]$	27/1263/29	1.89
MBPP	500	34.6	$-1.4\,[-3.0, +0.2]$	5/483/12	1.81
MATH500	500	58.8	$-0.2\,[-2.2, +2.0]$	14/471/15	1.76
SDAR-8B					
HumanEval	164	43.9	$-2.4\,[-6.1, +1.2]$	3/154/7	1.38
GSM8K	1319	50.5	$+0.3\,[-0.5, +1.1]$	18/1287/14	1.94
MBPP	500	33.8	$+0.6\,[-1.6, +2.8]$	16/471/13	1.84
MATH500	500	64.6	$-1.4\,[-3.4, +0.6]$	10/473/17	1.86

C Additional details on numerical identity

Frozen-state numerical audit. Frozen-state comparisons use 1,024 states per model. Replica batches contain copies of the same state, whereas mixed batches combine equal-length states from different decoding steps of the same prompt; neither setting represents general mixed-length serving. The float32 condition recomputes logits from the canonical hidden states through the language-model head, including Dream's positional shift, rather than evaluating the full backbone in float32. Repeating the canonical configuration and enabling the tested deterministic mode are both bit-exact. Under the other configurations, selection changes concentrate on positions with nearly tied scores.

Interpretation of uncertainty. For LLaDA-1.5 and Dream-Instruct, respectively, the paired accuracy-interval half-widths for the replica-batch comparison are 2.44/3.05 points on HumanEval, 0.91/1.29 on GSM8K, 1.60/1.60 on MBPP, and 2.20/2.00 on MATH500. These values summarize sampling precision for this intervention; they are not equivalence margins or confidence bounds on an intrinsic numerical noise process. An interval containing zero therefore leaves the accuracy effect unresolved rather than establishing equivalence.

Full-generation replica batching. Table 4 reports the raw token-array mismatch, paired accuracy difference, directional correctness changes, and retained-item runtime for canonical batch-one and replica batch-two decoding across all five models and four benchmarks. Raw mismatch rates range from 24.0% to 88.9%, while every paired accuracy interval contains zero. Replica execution requires 1.38–1.94 times the canonical per-item runtime because it intentionally duplicates computation for one retained output.

Layered divergence audit. The layered audit applies four comparisons to the same canonical and replica-batch output pairs reported in Table 4. The *raw-window* layer compares the complete retained

Table 5: **Selector effects across numerical configurations.** Each Δ is the selector's paired accuracy difference from confidence within one configuration. $I(\text{numerics})$ is the jointly resampled difference between those effects. Intervals are nominal, unadjusted 95% confidence intervals; a star marks an interval excluding zero. Flip columns report correctness changes between configurations.

Task	Arm	Δ canonical	Δ batch two	$I(\text{numerics})$	Conf. flips	Arm flips
LLaDA-1.5						
GSM8K	DAPD	$+3.5\,[+1.3, +5.7]^*$	$+4.2\,[+2.0, +6.4]^*$	$+0.8\,[-1.7, +3.3]$	15%	14%
HumanEval	Margin	$+6.1\,[+0.6, +11.6]^*$	$+4.3\,[-1.8, +11.0]$	$-1.8\,[-7.9, +4.9]$	5%	12%
MBPP	DAPD	$+3.8\,[+1.6, +6.0]^*$	$+5.6\,[+3.0, +8.4]^*$	$+1.8\,[-0.4, +4.0]$	5%	5%
MBPP	ADAS	$+5.8\,[+2.2, +9.4]^*$	$+4.6\,[+1.0, +8.4]^*$	$-1.2\,[-4.6, +2.2]$	5%	9%
MBPP	Margin	$+3.6\,[+0.4, +6.8]^*$	$+6.4\,[+3.2, +9.6]^*$	$+2.8\,[-0.2, +5.8]$	5%	6%
Dream-Instruct						
GSM8K	DAPD	$+2.3\,[+0.1, +4.5]^*$	$+0.5\,[-1.7, +2.6]$	$-1.8\,[-4.6, +0.9]$	20%	19%

integer arrays. The *pre-termination* layer truncates each array at its own first end-of-sequence or end-of-turn token; outputs without such a token retain their full window. The *scorer-input* layer compares the normalized text or program received by the task scorer, including stop-sequence truncation for the mathematical tasks and fence and signature handling for the code tasks. Finally, the *answer or outcome* layer compares normalized numeric strings on GSM8K, Math-Verify matches on MATH500, and execution outcomes on HumanEval and MBPP. Different programs with the same test outcome are not assumed to be semantically equivalent.

For MATH500, answers are parsed and compared using `math_verify` 0.9.0. Because its comparison can depend on argument order, both orders are evaluated and an answer pair is considered matched if either order succeeds. Both `parse()` and `verify()` are called with `raise_on_error=True`, allowing internal failures to be distinguished from genuine non-matches. No internal exception occurs in any of the 20 cells. The two argument orders disagree on one of the 495 both-extracted Dream-Instruct MATH500 items and on none of the other models. A non-match therefore means that the specified Math-Verify procedure finds no match in either direction; it is not a proof of mathematical inequivalence.

Every model–task cell contains the expected 164, 1,319, 500, or 500 items, with identical canonical and replica-batch index sets and no duplicate item indices. Recomputing the raw mismatch rates and paired accuracy differences reproduces Table 4 to rounding in all 20 cells. Independent scoring differs from the stored correctness label on one item in each of three cells—LLaDA-1.5 HumanEval, LLaDA-1.5 GSM8K, and Dream-Instruct HumanEval—and agrees in every other cell. Raw-array disagreement is unaffected.

Extraction failures are retained as a separate, unclassified outcome and are never counted as answer agreement. At least one side fails extraction for 5 Dream-Instruct GSM8K items, 5 Dream-Instruct MATH500 items, 1 LLaDA-8B-Instruct MATH500 item, 1 Dream-Base GSM8K item, 14 Dream-Base MATH500 items, and 4 SDAR-8B MATH500 items; the other mathematical cells contain none. Both sides fail on 5, 3, 0, 0, 9, and 4 of these items, respectively.

A gain denotes an item that is incorrect under canonical execution and correct under replica-batch execution; a loss denotes the reverse. Both occur in every audited setting and partly offset in aggregate. For example, LLaDA-1.5/GSM8K contains 13 gains and 22 losses, producing a net change of only -0.7 points despite 35 item-level correctness changes. Small aggregate differences therefore do not imply agreement on which items are solved.

Selection–numerics interaction. For the six selector comparisons available under both canonical and replica-batch execution, Table 5 reports each selector's effect relative to confidence within each configuration and their jointly resampled difference, $I(\text{numerics})$. The interaction uses 10,000 shared bootstrap draws across the four relevant outcome vectors. None of the six nominal 95% intervals excludes zero, so no change in selector advantage is detected at the available sample sizes. This does not establish equality or practical equivalence; for example, the LLaDA-1.5/MBPP margin interaction remains compatible with changes from -0.2 to $+5.8$ points.

Table 6: **SDAR $k = 4$ selector effects under three evaluation conditions.** Values are selector-minus-confidence accuracy differences in percentage points. Conditions A, B, and C respectively use the associated scorer on retained outputs, the unified scorer on the same outputs, and the unified scorer on tie-controlled outputs. Intervals are paired nominal 95% bootstrap intervals; a star marks an interval excluding zero.

Selector	Task	Condition A	Condition B	Condition C
DAPD	HumanEval	$-5.49\,[-10.37, -0.61]^*$	$-5.49\,[-10.37, -0.61]^*$	$-5.49\,[-10.37, -0.61]^*$
DAPD	GSM8K	$-6.82\,[-10.24, -3.41]^*$	$+3.26\,[+0.53, +5.99]^*$	$+3.26\,[+0.53, +5.99]^*$
DAPD	MBPP	$+0.60\,[-2.40, +3.60]$	$+0.60\,[-2.40, +3.60]$	$+0.60\,[-2.40, +3.60]$
DAPD	MATH500	$-2.80\,[-6.60, +0.80]$	$-2.80\,[-6.80, +1.00]$	$-2.80\,[-6.80, +1.00]$
ADAS	HumanEval	$-10.37\,[-15.85, -4.88]^*$	$-10.37\,[-16.46, -4.88]^*$	$-10.37\,[-16.46, -4.88]^*$
ADAS	GSM8K	$-0.23\,[-3.49, +3.18]$	$+0.38\,[-2.43, +3.26]$	$+0.38\,[-2.43, +3.26]$
ADAS	MBPP	$+2.40\,[-1.00, +5.80]$	$+2.40\,[-0.80, +5.80]$	$+2.40\,[-0.80, +5.80]$
ADAS	MATH500	$-0.40\,[-4.20, +3.60]$	$-0.40\,[-4.20, +3.40]$	$-0.40\,[-4.20, +3.40]$
Margin	HumanEval	$-2.44\,[-9.15, +3.66]$	$-2.44\,[-8.54, +3.66]$	$-2.44\,[-9.15, +4.27]$
Margin	GSM8K	$-4.70\,[-7.96, -1.36]^*$	$-1.14\,[-3.94, +1.59]$	$+0.53\,[-2.27, +3.26]$
Margin	MBPP	$+0.00\,[-3.00, +3.00]$	$+0.00\,[-3.00, +3.00]$	$-0.80\,[-3.80, +2.20]$
Margin	MATH500	$-4.20\,[-8.20, -0.20]^*$	$-4.20\,[-8.20, -0.20]^*$	$-0.80\,[-4.80, +3.20]$

Table 7: **Decomposition of SDAR effect changes.** Contributions to the condition-A-to-C change are computed from item counts before rounding and reported in percentage points. Omitted cells are zero in every column. A zero net contribution does not imply identical generated text; margin/HumanEval, for example, differs on 50 of 164 completions.

Selector	Task	Sel. scorer	Sel. generation	Conf. scorer	Conf. generation	Effect Δ
DAPD	GSM8K	$+32.30$	$+0.00$	$+22.21$	$+0.00$	$+10.08$
ADAS	GSM8K	$+22.82$	$+0.00$	$+22.21$	$+0.00$	$+0.61$
Margin	GSM8K	$+25.78$	$+1.67$	$+22.21$	$+0.00$	$+5.23$
Margin	MBPP	$+0.00$	-0.80	$+0.00$	$+0.00$	-0.80
Margin	MATH500	$+0.00$	$+3.40$	$+0.00$	$+0.00$	$+3.40$

Scorer and generation sensitivity on SDAR. The three SDAR evaluation conditions defined in Appendix A.2 separate scorer sensitivity from generation sensitivity: condition A applies the associated scorer to the retained outputs, condition B applies the unified scorer to the same outputs, and condition C applies the unified scorer to outputs generated under the explicit lowest-index tie convention. Table 6 reports selector effects under all three conditions, and Table 7 decomposes each A-to-C change into selector-side and confidence-side scoring and generation contributions.

Intervals for conditions B and C are computed from their respective item-level vectors using the same bootstrap draws. DAPD and ADAS have identical selector outputs in B and C across all eight task cells. Confidence is also identical in six of those cells; on GSM8K, two of its correctness labels change but offset in aggregate. Identical intervals are therefore not interpreted as evidence that the underlying vectors are identical.

The largest sensitivity occurs for DAPD on GSM8K. Its measured effect changes from -6.82 points under condition A to $+3.26$ points under conditions B and C, although DAPD's generated text is unchanged. DAPD's scorer contribution is $+32.30$ points and confidence's is $+22.21$ points, yielding the $+10.08$-point change in their paired effect. ADAS's $+0.61$-point GSM8K effect change is likewise entirely attributable to scoring. Margin additionally has generation contributions of $+1.67$ points on GSM8K, -0.80 on MBPP, and $+3.40$ on MATH500.

Table 8: **Fixed-pace and adaptive decoding.** Cells report accuracy (%) with mean model evaluations in parentheses. Fixed paces use confidence selection for LLaDA and SDAR and entropy selection for Dream. Dream's adaptive comparison therefore changes both selection evidence and pace.

Task	$k = 1$	$k = 2$	$k = 4$	$k = 8$	Adaptive
LLaDA-1.5					
HumanEval	43.9 (256)	36.6 (128)	22.0 (64)	10.4 (32)	43.3 (101)
GSM8K	80.5 (256)	78.4 (128)	68.0 (64)	39.0 (32)	81.0 (75)
MBPP	38.0 (256)	31.0 (128)	20.6 (64)	10.4 (32)	38.0 (42)
MATH500	33.6 (256)	29.0 (128)	25.6 (64)	16.2 (32)	32.8 (97)
Dream-Instruct					
HumanEval	53.7 (256)	26.2 (128)	11.6 (64)	5.5 (32)	57.9 (91)
GSM8K	76.3 (256)	68.0 (128)	50.0 (64)	22.6 (32)	74.8 (62)
MBPP	54.8 (256)	37.6 (128)	20.2 (64)	10.4 (32)	55.4 (63)
MATH500	37.0 (256)	32.6 (128)	20.6 (64)	5.0 (32)	37.4 (115)
LLaDA-8B-Instruct					
HumanEval	40.9 (256)	27.4 (128)	15.9 (64)	3.7 (32)	41.5 (78)
GSM8K	78.0 (256)	73.5 (128)	63.4 (64)	31.5 (32)	78.1 (78)
MBPP	36.4 (256)	24.2 (128)	12.8 (64)	4.8 (32)	36.4 (70)
MATH500	31.6 (256)	26.6 (128)	19.6 (64)	7.4 (32)	33.2 (98)
Dream-Base					
HumanEval	42.1 (256)	26.8 (128)	14.6 (64)	4.9 (32)	43.3 (135)
GSM8K	75.1 (256)	64.5 (128)	40.6 (64)	4.7 (32)	74.5 (152)
MBPP	52.8 (256)	36.8 (128)	19.0 (64)	9.8 (32)	53.6 (91)
MATH500	33.6 (256)	29.2 (128)	18.0 (64)	2.6 (32)	32.2 (102)
SDAR-8B					
HumanEval	62.2 (256)	45.7 (128)	25.6 (64)	9.1 (32)	62.2 (64)
GSM8K	88.3 (256)	84.5 (128)	70.7 (64)	28.6 (32)	86.7 (61)
MBPP	45.6 (256)	22.0 (128)	11.6 (64)	3.0 (32)	43.6 (41)
MATH500	43.2 (256)	37.2 (128)	24.8 (64)	7.2 (32)	42.8 (92)

D Additional details on pace and selection

Fixed-selector pace comparisons. Table 8 reports accuracy and mean model evaluations at $k = 1, 2, 4, 8$ and under each model's adaptive confidence-threshold schedule. Table 9 gives the paired accuracy change for all 60 adjacent fixed-pace transitions. Accuracy decreases in every transition, with 58 of the 60 nominal 95% confidence intervals excluding zero. Table 10 compares adaptive confidence thresholding directly with $k = 1$. The adaptive schedule uses 16–60% of the evaluations required by $k = 1$. None of the intervals excludes zero on the four LLaDA and Dream models, including the 5,000-item MATH comparisons. On SDAR-8B, the adaptive schedule is 1.6 points lower on GSM8K and 2.0 points lower on MBPP.

Table 9: **Accuracy changes between adjacent fixed paces.** Values are paired percentage-point changes with nominal 95% bootstrap confidence intervals using 10,000 resamples. A star marks an interval excluding zero.

Task	$k = 1 \to 2$	$k = 2 \to 4$	$k = 4 \to 8$
LLaDA-1.5			
HumanEval	$-7.3\,[-12.8, -1.8]^*$	$-14.6\,[-21.3, -8.5]^*$	$-11.6\,[-17.7, -5.5]^*$
GSM8K	$-2.1\,[-4.2, +0.0]$	$-10.4\,[-13.0, -7.7]^*$	$-29.0\,[-32.1, -25.8]^*$
MBPP	$-7.0\,[-10.4, -3.6]^*$	$-10.4\,[-13.8, -7.0]^*$	$-10.2\,[-13.4, -7.0]^*$
MATH500	$-4.6\,[-8.0, -1.2]^*$	$-3.4\,[-7.2, +0.4]$	$-9.4\,[-13.4, -5.6]^*$
Dream-Instruct			
HumanEval	$-27.4\,[-35.4, -19.5]^*$	$-14.6\,[-22.0, -7.3]^*$	$-6.1\,[-11.0, -1.2]^*$
GSM8K	$-8.3\,[-10.8, -5.8]^*$	$-18.0\,[-21.0, -15.1]^*$	$-27.4\,[-30.6, -24.3]^*$
MBPP	$-17.2\,[-21.2, -13.4]^*$	$-17.4\,[-21.6, -13.4]^*$	$-9.8\,[-13.2, -6.4]^*$
MATH500	$-4.4\,[-8.0, -1.0]^*$	$-12.0\,[-16.0, -8.0]^*$	$-15.6\,[-19.4, -12.0]^*$
LLaDA-8B-Instruct			
HumanEval	$-13.4\,[-20.1, -6.7]^*$	$-11.6\,[-17.7, -5.5]^*$	$-12.2\,[-17.7, -7.3]^*$
GSM8K	$-4.5\,[-6.7, -2.3]^*$	$-10.2\,[-13.0, -7.4]^*$	$-31.9\,[-35.1, -28.7]^*$
MBPP	$-12.2\,[-16.0, -8.4]^*$	$-11.4\,[-15.0, -7.8]^*$	$-8.0\,[-11.0, -5.0]^*$
MATH500	$-5.0\,[-8.4, -1.6]^*$	$-7.0\,[-10.8, -3.2]^*$	$-12.2\,[-16.0, -8.6]^*$
Dream-Base			
HumanEval	$-15.2\,[-22.0, -8.5]^*$	$-12.2\,[-18.3, -6.1]^*$	$-9.8\,[-14.6, -4.9]^*$
GSM8K	$-10.6\,[-13.1, -8.1]^*$	$-24.0\,[-27.0, -20.9]^*$	$-35.9\,[-38.6, -33.2]^*$
MBPP	$-16.0\,[-20.0, -12.2]^*$	$-17.8\,[-22.0, -13.6]^*$	$-9.2\,[-12.6, -5.8]^*$
MATH500	$-4.4\,[-7.8, -1.0]^*$	$-11.2\,[-14.8, -7.6]^*$	$-15.4\,[-18.8, -12.0]^*$
SDAR-8B			
HumanEval	$-16.5\,[-23.8, -9.1]^*$	$-20.1\,[-28.0, -12.8]^*$	$-16.5\,[-22.6, -11.0]^*$
GSM8K	$-3.9\,[-5.5, -2.3]^*$	$-13.7\,[-16.2, -11.3]^*$	$-42.2\,[-45.2, -39.1]^*$
MBPP	$-23.6\,[-28.0, -19.4]^*$	$-10.4\,[-14.2, -6.6]^*$	$-8.6\,[-11.4, -5.8]^*$
MATH500	$-6.0\,[-9.8, -2.2]^*$	$-12.4\,[-16.4, -8.4]^*$	$-17.6\,[-21.6, -13.8]^*$

Table 10: **Adaptive confidence thresholding relative to $k = 1$.** Values are paired accuracy differences in percentage points. W/T/L denotes adaptive-only correct, unchanged correctness, and $k = 1$-only correct. The evaluation ratio is adaptive evaluations divided by $k = 1$ evaluations. Dream uses different selection signals in the two policies.

Task	Δ accuracy [95% CI]	W/T/L	Eval. ratio
LLaDA-1.5			
HumanEval	$-0.6\,[-1.8, +0.0]$	0/163/1	0.39
GSM8K	$+0.5\,[-0.3, +1.4]$	20/1286/13	0.29
MBPP	$+0.0\,[-0.6, +0.6]$	1/498/1	0.17
MATH500	$-0.8\,[-2.0, +0.4]$	3/490/7	0.38
MATH ($n = 5{,}000$)	$+0.3\,[-0.1, +0.8]$	72/4873/55	0.37
Dream-Instruct			
HumanEval	$+4.3\,[-2.4, +11.0]$	19/133/12	0.35
GSM8K	$-1.5\,[-3.8, +0.7]$	104/1091/124	0.24
MBPP	$+0.6\,[-2.8, +4.0]$	40/423/37	0.25
MATH500	$+0.4\,[-1.8, +2.6]$	16/470/14	0.45
MATH ($n = 5{,}000$)	$+0.0\,[-1.1, +1.2]$	409/4184/407	0.45
LLaDA-8B-Instruct			
HumanEval	$+0.6\,[-1.2, +3.0]$	2/161/1	0.30
GSM8K	$+0.1\,[-0.6, +0.8]$	12/1296/11	0.31
MBPP	$+0.0\,[-0.6, +0.6]$	1/498/1	0.27
MATH500	$+1.6\,[-1.8, +5.0]$	43/422/35	0.38
Dream-Base			
HumanEval	$+1.2\,[-4.9, +7.3]$	13/140/11	0.53
GSM8K	$-0.6\,[-2.7, +1.4]$	88/1135/96	0.60
MBPP	$+0.8\,[-2.6, +4.2]$	40/424/36	0.35
MATH500	$-1.4\,[-4.0, +1.2]$	17/459/24	0.40
SDAR-8B			
HumanEval	$+0.0\,[-2.4, +2.4]$	2/160/2	0.25
GSM8K	$-1.6\,[-2.4, -0.8]$	4/1290/25	0.24
MBPP	$-2.0\,[-3.4, -0.8]$	1/488/11	0.16
MATH500	$-0.4\,[-1.8, +1.0]$	5/488/7	0.36

Table 11: **Selection–pace interactions on GSM8K.** Each model contributes confidence, adapted DAPD, and adapted ADAS at $k = 4$ and $k = 8$ under a common harness and scorer. I_s is the jointly resampled interaction and $J = I_{\mathrm{ADAS}} - I_{\mathrm{DAPD}}$. Values are percentage points with nominal 95% confidence intervals; a star marks an interval excluding zero.

Model	I_{ADAS}	I_{DAPD}	J
LLaDA-1.5	$+6.8\,[+2.5, +11.1]^*$	$-1.5\,[-4.2, +1.3]$	$+8.3\,[+4.2, +12.4]^*$
LLaDA-8B-Instruct	$+6.6\,[+2.2, +11.0]^*$	$-3.4\,[-6.4, -0.4]^*$	$+10.0\,[+5.6, +14.5]^*$
Dream-Base	$+6.8\,[+2.9, +10.8]^*$	$-3.4\,[-5.6, -1.3]^*$	$+10.2\,[+6.1, +14.3]^*$
SDAR-8B	$+14.8\,[+10.7, +19.0]^*$	$+7.1\,[+3.3, +10.9]^*$	$+7.7\,[+3.5, +11.9]^*$
Dream-Instruct	$+1.4\,[-2.9, +5.6]$	$-0.5\,[-3.0, +2.1]$	$+1.9\,[-2.4, +6.1]$

Selection–pace interactions. We evaluate confidence, adapted DAPD, and adapted ADAS at $k = 4$ and $k = 8$ using a common harness and scorer within each model–task crossing. All six arms contain the complete benchmark item set, and the paired bootstrap jointly resamples the same items across the six arms using 10,000 draws. Table 11 reports the interactions I_{ADAS}, I_{DAPD}, and $J = I_{\mathrm{ADAS}} - I_{\mathrm{DAPD}}$ defined in Equation 1. Table 12 provides the six underlying accuracies and pace-specific effects.

Across the five GSM8K crossings, I_{ADAS} is positive on every model and its interval excludes zero on four. I_{DAPD} is negative on LLaDA-8B-Instruct and Dream-Base, positive on SDAR-8B, and unresolved on LLaDA-1.5 and Dream-Instruct. The direct contrast J is positive on all five models and excludes zero on four, where it ranges from 7.7 to 10.2 percentage points. These five crossings use the same 1,319-item GSM8K benchmark rather than independently sampled task panels.

The LLaDA-1.5/MBPP crossing does not show the same pattern: $I_{\mathrm{ADAS}} = +0.2$, $I_{\mathrm{DAPD}} = -2.0$, and $J = +2.2$, with all three intervals containing zero. The result therefore supports cross-model consistency on GSM8K but does not establish cross-task generality.

Exploratory MBPP interactions. Table 13 reports four additional MBPP interactions assembled from separately generated $k = 4$ and $k = 8$ arms. These comparisons use the same jointly resampled interaction estimator but do not share the controlled six-arm generation procedure used above. None resolves an ADAS interaction. LLaDA-8B-Instruct has a nominally negative DAPD interaction of -3.6 points with an interval excluding zero, but the comparison is exploratory and unadjusted for multiplicity.

Table 12: **Six-arm accounting for the selection–pace interactions.** Each crossing evaluates confidence, adapted DAPD, and adapted ADAS at $k = 4$ and $k = 8$. Effects are paired selector-minus-confidence accuracy differences with nominal 95% bootstrap intervals using 10,000 resamples. GSM8K uses $n = 1{,}319$ and MBPP uses $n = 500$.

Arm	Acc. $k = 4$	Acc. $k = 8$	Effect $k = 4$	Effect $k = 8$
LLaDA-1.5 / GSM8K				
Confidence	68.01	39.80	–	–
DAPD	71.72	42.00	$+3.71\,[+1.52, +5.91]$	$+2.20\,[+0.30, +4.09]$
ADAS	67.40	46.02	$-0.61\,[-3.41, +2.20]$	$+6.22\,[+2.88, +9.55]$
LLaDA-8B-Instruct / GSM8K				
Confidence	63.91	31.54	–	–
DAPD	67.55	31.77	$+3.64\,[+1.21, +5.99]$	$+0.23\,[-1.52, +1.97]$
ADAS	65.12	39.35	$+1.21\,[-1.74, +4.32]$	$+7.81\,[+4.55, +11.07]$
Dream-Base / GSM8K				
Confidence	50.57	12.89	–	–
DAPD	53.90	12.81	$+3.34\,[+1.36, +5.38]$	$-0.08\,[-0.68, +0.53]$
ADAS	51.63	20.77	$+1.06\,[-2.12, +4.25]$	$+7.88\,[+5.38, +10.39]$
SDAR-8B / GSM8K				
Confidence	70.73	28.58	–	–
DAPD	74.00	38.89	$+3.26\,[+0.61, +5.99]$	$+10.31\,[+7.43, +13.12]$
ADAS	71.11	43.74	$+0.38\,[-2.50, +3.18]$	$+15.16\,[+12.06, +18.35]$
Dream-Instruct / GSM8K				
Confidence	53.68	24.34	–	–
DAPD	55.80	25.93	$+2.12\,[-0.23, +4.40]$	$+1.59\,[+0.30, +2.96]$
ADAS	55.80	27.82	$+2.12\,[-0.91, +5.16]$	$+3.49\,[+0.68, +6.44]$
LLaDA-1.5 / MBPP				
Confidence	20.60	10.40	–	–
DAPD	24.40	12.20	$+3.80\,[+1.60, +6.20]$	$+1.80\,[+0.60, +3.20]$
ADAS	26.40	16.40	$+5.80\,[+2.20, +9.40]$	$+6.00\,[+2.60, +9.40]$

Table 13: **Exploratory selection–pace interactions on MBPP.** The $k = 4$ and $k = 8$ arms come from separate evaluation sets rather than a common six-arm generation procedure. Values are percentage points with nominal, unadjusted 95% bootstrap intervals using shared item resamples. A star marks an interval excluding zero.

Model	I_{ADAS}	I_{DAPD}	J
Dream-Instruct	$-1.2\,[-6.2, +4.0]$	$+1.6\,[-1.2, +4.4]$	$-2.8\,[-8.0, +2.4]$
LLaDA-8B-Instruct	$-1.8\,[-6.0, +2.4]$	$-3.6\,[-7.0, -0.2]^{*}$	$+1.8\,[-2.6, +6.4]$
Dream-Base	$-0.8\,[-6.0, +4.2]$	$-0.6\,[-3.0, +1.8]$	$-0.2\,[-5.4, +5.0]$
SDAR-8B	$-1.0\,[-5.0, +3.0]$	$-1.0\,[-4.4, +2.4]$	$-0.0\,[-3.8, +3.8]$

Table 14: **Additional project-harness outcomes at $k = 8$.** Values are confidence accuracy, tested-policy accuracy, and their paired difference in percentage points. DAPD and ADAS are fixed-pace adaptations. GSM8K DAPD and ADAS results appear in Table 12.

Task	Arm	Confidence	Arm accuracy	Δ [95% CI]
LLaDA-1.5				
GSM8K	Margin	40.0	43.4	$+3.4\,[+0.2, +6.7]^*$
HumanEval	DAPD	10.4	7.3	$-3.0\,[-6.1, +0.0]$
HumanEval	ADAS	10.4	9.8	$-0.6\,[-4.9, +3.7]$
HumanEval	Margin	10.4	12.8	$+2.4\,[-2.4, +7.3]$
MATH500	DAPD	13.4	13.6	$+0.2\,[-2.0, +2.4]$
MATH500	ADAS	13.4	15.8	$+2.4\,[-0.8, +5.8]$
MATH500	Margin	13.4	14.4	$+1.0\,[-2.2, +4.2]$
MBPP	DAPD	10.4	11.0	$+0.6\,[-1.2, +2.4]$
MBPP	ADAS	10.4	16.2	$+5.8\,[+2.6, +9.2]^*$
MBPP	Margin	10.4	15.0	$+4.6\,[+1.4, +7.8]^*$
Dream-Instruct				
GSM8K	Margin	24.6	23.4	$-1.1\,[-3.8, +1.4]$
HumanEval	DAPD	6.1	6.7	$+0.6\,[+0.0, +1.8]$
HumanEval	ADAS	6.1	5.5	$-0.6\,[-4.9, +3.7]$
HumanEval	Margin	6.1	10.4	$+4.3\,[-0.6, +9.1]$
MATH500	DAPD	5.0	5.6	$+0.6\,[+0.0, +1.4]$
MATH500	ADAS	5.0	7.2	$+2.2\,[-0.6, +5.0]$
MATH500	Margin	5.0	4.8	$-0.2\,[-2.6, +2.2]$
MBPP	DAPD	13.6	14.4	$+0.8\,[-0.2, +2.0]$
MBPP	ADAS	13.6	14.8	$+1.2\,[-2.2, +4.6]$
MBPP	Margin	13.6	13.0	$-0.6\,[-4.0, +3.0]$

Additional $k = 8$ comparisons. Table 14 reports the remaining $k = 8$ project-harness comparisons on LLaDA-1.5 and Dream-Instruct. DAPD and ADAS on GSM8K are already included in Table 12 and are not repeated. Table 15 pairs each listed $k = 4$ effect with the absolute accuracy change of confidence and the tested policy from $k = 4$ to $k = 8$. Every listed policy loses absolute accuracy at the faster pace, reinforcing that the interactions above describe changes in relative advantage rather than absolute improvements.

Table 16 characterizes the adapted structural selectors at $k = 8$. DAPD proposes fewer than eight positions on 56–99% of steps and overlaps confidence's top-eight set by 0.79–0.97. ADAS always proposes eight positions in these runs and overlaps confidence by 0.70–0.85. Consequently, increasing k changes not only update size but also each adapted policy's reliance on confidence padding.

Table 15: **Fixed-pace effects and within-policy accuracy changes.** The $k = 4$ column reports each arm's paired effect relative to confidence. The final columns report each policy's absolute accuracy change from $k = 4$ to $k = 8$. Separate pace-specific intervals do not constitute an interaction test.

Task	Arm	Δ at $k = 4$	Confidence $4 \to 8$	Arm $4 \to 8$
LLaDA-1.5				
GSM8K	Margin	$+1.5\,[-1.3, +4.3]$	$-27.7\,[-30.9, -24.6]$	$-25.9\,[-29.0, -22.7]$
HumanEval	DAPD	$+0.0\,[-3.7, +3.7]$	$-11.6\,[-17.7, -5.5]$	$-14.6\,[-20.7, -9.1]$
HumanEval	ADAS	$-0.6\,[-5.5, +4.9]$	$-11.6\,[-17.7, -5.5]$	$-11.6\,[-17.1, -6.7]$
HumanEval	Margin	$+6.1\,[+0.6, +11.6]^*$	$-11.6\,[-17.7, -5.5]$	$-15.2\,[-22.0, -9.1]$
MATH500	DAPD	$+0.2\,[-2.6, +3.2]$	$-12.2\,[-16.0, -8.4]$	$-12.2\,[-16.2, -8.2]$
MATH500	ADAS	$-1.0\,[-4.6, +2.6]$	$-12.2\,[-16.0, -8.4]$	$-8.8\,[-12.6, -5.2]$
MATH500	Margin	$-1.0\,[-4.4, +2.6]$	$-12.2\,[-16.0, -8.4]$	$-10.2\,[-14.0, -6.6]$
MBPP	DAPD	$+3.8\,[+1.6, +6.0]^*$	$-10.2\,[-13.4, -7.0]$	$-13.4\,[-16.8, -10.0]$
MBPP	ADAS	$+5.8\,[+2.2, +9.4]^*$	$-10.2\,[-13.4, -7.0]$	$-10.2\,[-14.0, -6.4]$
MBPP	Margin	$+3.6\,[+0.4, +6.8]^*$	$-10.2\,[-13.4, -7.0]$	$-9.2\,[-12.8, -5.8]$
Dream-Instruct				
GSM8K	Margin	$-2.5\,[-5.6, +0.6]$	$-29.6\,[-32.6, -26.7]$	$-28.3\,[-31.5, -25.0]$
HumanEval	DAPD	$-0.6\,[-3.7, +2.4]$	$-14.0\,[-19.5, -8.5]$	$-12.8\,[-18.3, -7.9]$
HumanEval	ADAS	$+0.6\,[-6.7, +7.9]$	$-14.0\,[-19.5, -8.5]$	$-15.2\,[-21.3, -9.8]$
HumanEval	Margin	$-1.8\,[-7.9, +4.3]$	$-14.0\,[-19.5, -8.5]$	$-7.9\,[-13.4, -2.4]$
MATH500	DAPD	$+0.6\,[-2.0, +3.2]$	$-15.6\,[-19.4, -12.0]$	$-15.6\,[-19.4, -11.8]$
MATH500	ADAS	$+0.2\,[-3.8, +4.2]$	$-15.6\,[-19.4, -12.0]$	$-13.6\,[-17.4, -9.8]$
MATH500	Margin	$-2.6\,[-6.2, +1.0]$	$-15.6\,[-19.4, -12.0]$	$-13.2\,[-16.8, -9.8]$
MBPP	DAPD	$-0.8\,[-3.4, +1.8]$	$-15.2\,[-19.4, -11.2]$	$-13.6\,[-17.6, -9.6]$
MBPP	ADAS	$+2.4\,[-1.6, +6.4]$	$-15.2\,[-19.4, -11.2]$	$-16.4\,[-20.4, -12.4]$
MBPP	Margin	$-2.0\,[-6.4, +2.2]$	$-15.2\,[-19.4, -11.2]$	$-13.8\,[-18.0, -9.8]$

Table 16: **Live selection behavior at $k = 8$.** Entries report mean structural-set size before padding, the percentage of steps proposing fewer than eight positions, and overlap with confidence's top-eight set.

Task	Arm	Mean size	Short (%)	Confidence overlap
LLaDA-1.5				
GSM8K	DAPD	2.98	90	0.93
GSM8K	ADAS	8.00	0	0.78
HumanEval	DAPD	4.90	79	0.89
HumanEval	ADAS	8.00	0	0.76
MATH500	DAPD	2.00	95	0.95
MATH500	ADAS	8.00	0	0.78
MBPP	DAPD	9.14	56	0.79
MBPP	ADAS	8.00	0	0.70
Dream-Instruct				
GSM8K	DAPD	2.49	90	0.94
GSM8K	ADAS	8.00	0	0.83
HumanEval	DAPD	2.01	94	0.96
HumanEval	ADAS	8.00	0	0.83
MATH500	DAPD	1.24	99	0.97
MATH500	ADAS	8.00	0	0.82
MBPP	DAPD	5.53	70	0.91
MBPP	ADAS	8.00	0	0.85

Table 17: **Counterfactual decomposition of token persistence.** Direct visibility is a rate difference. Peer-support columns report the change in the tested proposal's logit margin after committing its peers, split by confidence-reference agreement. The final column compares the contextual effects of high-interaction and distance-matched random peers. Intervals are state-clustered 95% confidence intervals.

Model	Direct visibility	Peer support agreeing	Peer support disagreeing	High-interaction vs. random peer
LLaDA-1.5	$+0.071\,[0.062, 0.079]$	$+3.76\,[3.62, 3.91]$	$-1.33\,[-1.82, -0.84]$	$+2.17\,[2.00, 2.34]$
Dream-Instruct	$+0.035\,[0.026, 0.043]$	$+2.72\,[2.61, 2.83]$	$-2.87\,[-3.29, -2.42]$	$+1.49\,[1.35, 1.63]$
LLaDA-8B-Instruct	$+0.074\,[0.065, 0.083]$	$+3.69\,[3.54, 3.84]$	$-2.00\,[-2.43, -1.58]$	$+2.02\,[1.85, 2.19]$
Dream-Base	$+0.086\,[0.075, 0.097]$	$+2.26\,[2.16, 2.36]$	$-3.98\,[-4.30, -3.66]$	$+1.18\,[1.03, 1.33]$
SDAR-8B	$-0.45\,[-0.47, -0.43]$	$+12.5\,[11.6, 13.4]$	$-4.99\,[-7.21, -2.82]$	$-0.08\,[-0.67, 0.51]$

E Additional details on persistence

Counterfactual conditions. We separate a token's own visibility from contextual support using five counterfactual states. The *proposal state* keeps all candidate positions masked. The *committed state* makes the tested token and its selected peers visible. The *same-step leave-one-out state* keeps the peers visible but remasks the tested position, while the *next-step leave-one-out state* repeats this intervention after the next update. Finally, the *peer-substitution state* replaces either a high-interaction peer or a distance-matched random peer with its second-ranked token.

The difference between committed-state persistence and same-step leave-one-out recovery measures *direct visibility*. The change in the tested proposal's logit margin between the proposal and same-step leave-one-out states measures *contextual peer support*. These contrasts are necessary because a model that copies every visible token could exhibit perfect persistence without receiving any peer support at a masked position. The peer-substitution contrast tests whether the attention-based interaction score identifies peers with greater contextual influence.

Estimation and clustering. The analysis uses approximately 16 saved states from each of 64 prompts per model. Proposals are divided by whether they agree with the token selected by the sequential one-token confidence reference; this classification is not a ground-truth correctness label. Figure 4 (in the main manuscript) and Table 17 report the primary state-clustered estimates for direct visibility, peer support, and peer substitution across all five models.

Because states from the same prompt may be dependent, Table 18 additionally reports intervals obtained by resampling the 64 prompts. Point estimates are unchanged, while the intervals are generally wider. Nineteen of the 20 intervals exclude zero under both clustering choices; the high-interaction-versus-random contrast on SDAR-8B remains unresolved under both.

Outcome-stratified peer support. Table 19 stratifies peer support on LLaDA-1.5 and Dream-Instruct by the outcome obtained after committing the proposal and completing generation with confidence selection at $k = 4$. Correct/wrong rows distinguish the final rollout outcome, while helps/hurts/neutral rows compare that outcome with the corresponding baseline confidence rollout.

Table 18: **Prompt-clustered intervals for the persistence diagnostics.** Intervals are obtained by resampling 64 prompts per model with 10,000 bootstrap draws. Direct visibility is a rate difference; the remaining contrasts are measured in logits. A star marks an interval excluding zero.

Model	Direct visibility	Peer support agreeing	Peer support disagreeing	High-interaction vs. random peer
LLaDA-1.5	$[0.057, 0.085]^*$	$[3.44, 4.10]^*$	$[-2.14, -0.52]^*$	$[1.92, 2.40]^*$
Dream-Instruct	$[0.023, 0.046]^*$	$[2.46, 3.00]^*$	$[-3.52, -2.18]^*$	$[1.32, 1.66]^*$
LLaDA-8B-Instruct	$[0.061, 0.088]^*$	$[3.36, 4.04]^*$	$[-2.76, -1.27]^*$	$[1.79, 2.26]^*$
Dream-Base	$[0.068, 0.104]^*$	$[2.12, 2.40]^*$	$[-4.39, -3.53]^*$	$[1.01, 1.35]^*$
SDAR-8B	$[-0.56, -0.35]^*$	$[10.0, 15.1]^*$	$[-8.52, -2.13]^*$	$[-1.30, 1.10]$

Table 19: **Peer support stratified by rollout outcome.** Peer support is the change in proposal margin after committing peers, measured in logits with state-clustered 95% confidence intervals. Correct/wrong denotes final rollout correctness; helps/hurts/neutral denotes the correctness change relative to the baseline confidence rollout. A star marks an interval excluding zero.

Reference relation	Rollout stratum	Proposals	States	Peer support
LLaDA-1.5				
Agreeing	Correct	767	530	$+2.62\,[+2.23, +3.01]^*$
Agreeing	Helps	121	87	$+0.45\,[-0.39, +1.26]$
Agreeing	Hurts	91	65	$+2.31\,[+1.60, +3.06]^*$
Agreeing	Neutral	1504	1000	$+2.52\,[+2.24, +2.81]^*$
Agreeing	Wrong	949	622	$+2.16\,[+1.81, +2.50]^*$
Disagreeing	Correct	681	370	$-0.90\,[-1.69, -0.11]^*$
Disagreeing	Helps	146	85	$-0.21\,[-2.18, +1.73]$
Disagreeing	Hurts	77	52	$-1.45\,[-3.86, +1.19]$
Disagreeing	Neutral	1367	685	$-1.44\,[-1.96, -0.94]^*$
Disagreeing	Wrong	909	452	$-1.66\,[-2.27, -1.04]^*$
Dream-Instruct				
Agreeing	Correct	554	394	$+1.74\,[+1.37, +2.12]^*$
Agreeing	Helps	90	68	$+1.66\,[+0.59, +2.71]^*$
Agreeing	Hurts	79	52	$+1.90\,[+1.33, +2.48]^*$
Agreeing	Neutral	1614	1092	$+1.53\,[+1.33, +1.74]^*$
Agreeing	Wrong	1229	818	$+1.47\,[+1.24, +1.69]^*$
Disagreeing	Correct	453	252	$-2.09\,[-3.01, -1.13]^*$
Disagreeing	Helps	97	61	$-3.21\,[-4.76, -1.76]^*$
Disagreeing	Hurts	72	44	$-2.74\,[-4.28, -1.37]^*$
Disagreeing	Neutral	1348	693	$-2.85\,[-3.32, -2.38]^*$
Disagreeing	Wrong	1064	546	$-3.20\,[-3.67, -2.73]^*$

The direction of peer support follows reference agreement in every stratum. Reference-agreeing proposals retain positive support even when the completed answer is wrong, while reference-disagreeing proposals retain negative support even when the answer is correct. Reference agreement therefore describes support relative to the sequential confidence trajectory rather than token correctness or task utility. Because these strata condition on completed outcomes rather than intervening on correctness, they are descriptive and do not identify a causal effect of peer support on accuracy.

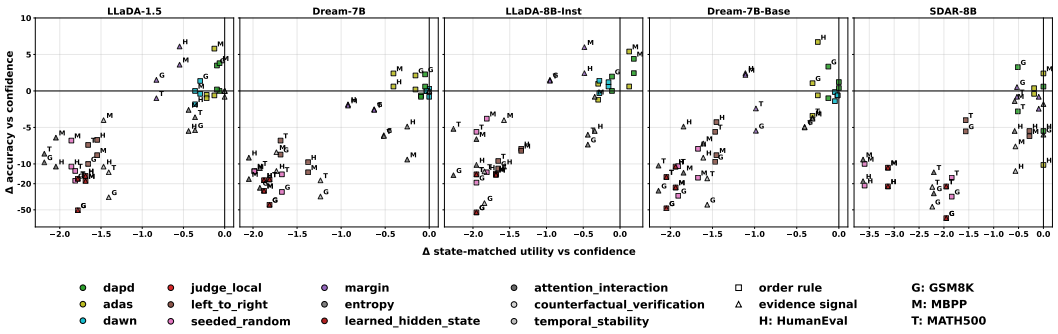

Figure 5: **Local reference agreement versus live policy performance.** The horizontal axis is accepted-depth difference from confidence, and the vertical axis is live accuracy difference at $k = 4$. Squares denote order rules and triangles evidence signals. Shaded regions provide a descriptive accuracy scale and are not equivalence bounds.

F Additional local-diagnostic and verification results

Fixed-pace policy comparisons. Tables 20 and 21 report confidence accuracy and the paired effect of each alternative order or evidence rule at $k = 4$ across five models and four tasks. The learned ranker appears as JUDGE in Table 20; its learned hidden-state evidence score uses the same computation and is therefore not repeated as an independent result in Table 21. The learned ranker is 11–67 percentage points below confidence in all 20 policy-matched comparisons.

For SDAR-8B, confidence, adapted DAPD, adapted ADAS, and margin use the tie-controlled evaluation conditions described in Appendix A.2. Other SDAR order and evidence policies are reported only as sensitivity analyses under their policy-specific tie handling and are marked with a dagger. DAWN is unavailable on SDAR-8B.

Reference agreement versus policy performance. Accepted depth measures agreement with a one-token confidence reference, not task correctness. Table 23 places this local metric beside the live policy effect for adapted DAPD, the learned ranker, and margin. Adapted DAPD and margin can lose accepted depth while improving live accuracy, whereas the learned ranker is negative on both measures across all models and tasks. Figure 5 extends the same comparison to every order and evidence rule.

The task-specific shaded regions in Figure 5 show the half-widths of the replica-batch accuracy intervals as visual scale references only. They are not equivalence regions or criteria for classifying policy effects.

One-action outcome diagnostics. Table 22 reports the outcome of replacing one confidence-selected action and then completing generation with confidence selection. Positive regret means that the substituted action lowers the final outcome relative to confidence. The analysis uses 128 frozen states from 16 prompts per model and reports prompt-clustered intervals.

Most intervals contain zero, and the diagnostic does not consistently separate policies that improve live performance from those that degrade it. SDAR-8B's development trajectories are almost never task-correct, so its outcome values and regrets provide little information about relative policy quality.

Verification fidelity and cost. Table 24 reports task accuracy, model evaluations, runtime, and output identity for the serial exact verifier VJUDGE and the separate FREEDAVE reproduction. VJUDGE reproduces the reference token arrays on every evaluated item but requires more model evaluations and longer runtime. FREEDAVE reduces both calls and runtime without a detected accuracy change, but does not reproduce every decoded output exactly. The two methods therefore establish different fidelity and efficiency properties and are not treated as interchangeable exactness claims.

Table 20: **Fixed-pace order comparisons at $k = 4$.** Confidence is absolute accuracy (%); other columns are paired percentage-point differences from confidence. A star marks a nominal 95% confidence interval excluding zero. A dagger marks an SDAR sensitivity result under policy-specific tie handling. LTR denotes left-to-right selection, and random uses a fixed seed.

Task	Conf.	DAPD	ADAS	DAWN	LTR	Random	Learned
LLaDA-1.5							
HumanEval	22.0	+0.0	−0.6	−1.8	−6.7*	−11.0*	−15.2*
GSM8K	67.8	+3.5*	−0.5	+1.4	−10.0*	−18.0*	−50.8*
MBPP	20.6	+3.8*	+5.8*	+0.0	−8.8*	−6.8*	−18.2*
MATH500	25.6	+0.2	−1.0	−0.4	−7.4*	−12.8*	−16.8*
Dream-Instruct							
HumanEval	20.1	−0.6	+0.6	+0.0	−9.8*	−12.8*	−17.7*
GSM8K	54.2	+2.3*	+2.1	+0.3	−8.7*	−26.7*	−41.8*
MBPP	28.8	−0.8	+2.4	+0.0	−13.4*	−13.0*	−25.6*
MATH500	20.6	+0.6	+0.2	−0.8	−6.8*	−14.4*	−17.2*
LLaDA-8B-Instruct							
HumanEval	21.3	+2.4	+0.6	+0.6	−7.9*	−13.4*	−14.0*
GSM8K	63.9	+2.0	+1.0	−0.3	−11.8*	−19.3*	−54.6*
MBPP	18.0	+4.4*	+5.4*	+1.2	−8.2*	−3.8*	−14.6*
MATH500	19.4	+0.0	−1.2	+1.4	−9.6*	−5.6*	−14.4*
Dream-Base							
HumanEval	11.0	+1.2	+6.7*	−0.6	−4.3*	−7.9*	−11.0*
GSM8K	50.6	+3.3*	+1.1	−0.2	−9.7*	−30.6*	−47.4*
MBPP	24.8	+0.4	−0.6	−0.6	−8.8*	−16.2*	−22.8*
MATH500	18.0	−1.0	−3.4	−1.4	−5.6*	−10.8*	−15.8*
SDAR-8B							
HumanEval	25.6	−5.5*	−10.4*	–	−5.5*†	−21.3*†	−22.0*†
GSM8K	70.7	+3.3*	+0.4	–	−5.5*†	−31.5*†	−66.9*†
MBPP	11.6	+0.6	+2.4	–	−6.2*†	−10.0*†	−11.4*†
MATH500	24.8	−2.8	−0.4	–	−4.0*†	−16.2*†	−22.0*†

Figure 7 relates wall-clock cost to model-call reduction. Every VJUDGE point remains above the runtime break-even line, whereas both FREEDAVE points fall below it. The comparison characterizes the evaluated implementations and workloads rather than all possible verification designs.

G LOCAL-RANKER AND VERIFIER CASE STUDIES

G.1 LEARNED MESSAGE-PASSING RANKER

We study the learned ranker JUDGE as an instrumented case study of whether pooled offline prediction transfers to token selection during generation. The candidate pool contains masked positions with confidence of at least 0.02. Candidate groups are formed from local spans of width two and four, hidden-state neighbors, code-bracket structure, and small random global groups.

Each position is represented by a 64-dimensional proposed-token embedding, six scalar features—confidence, log margin, entropy, mask ratio, position, and decoding time—and a 64-dimensional projection of the model's final hidden state. The hidden-state dimension is 4,096 for LLaDA and 3,584 for Dream. An initial unary logit is computed from these features. Three message-passing rounds transform member probabilities with a shared $1 \to 16 \to 16$ ReLU MLP, average messages from the other group members, and combine the result with the original unary logit to produce a position-level score.

Training uses weak group supervision. A group is labeled according to whether its realized quality change exceeds the training-set median, and the same binary label is assigned to every member. A position belonging to multiple groups contributes one loss term for each membership. Training minimizes binary cross-entropy using Adam with learning rate 3×10^{-4} for ten epochs, gradient

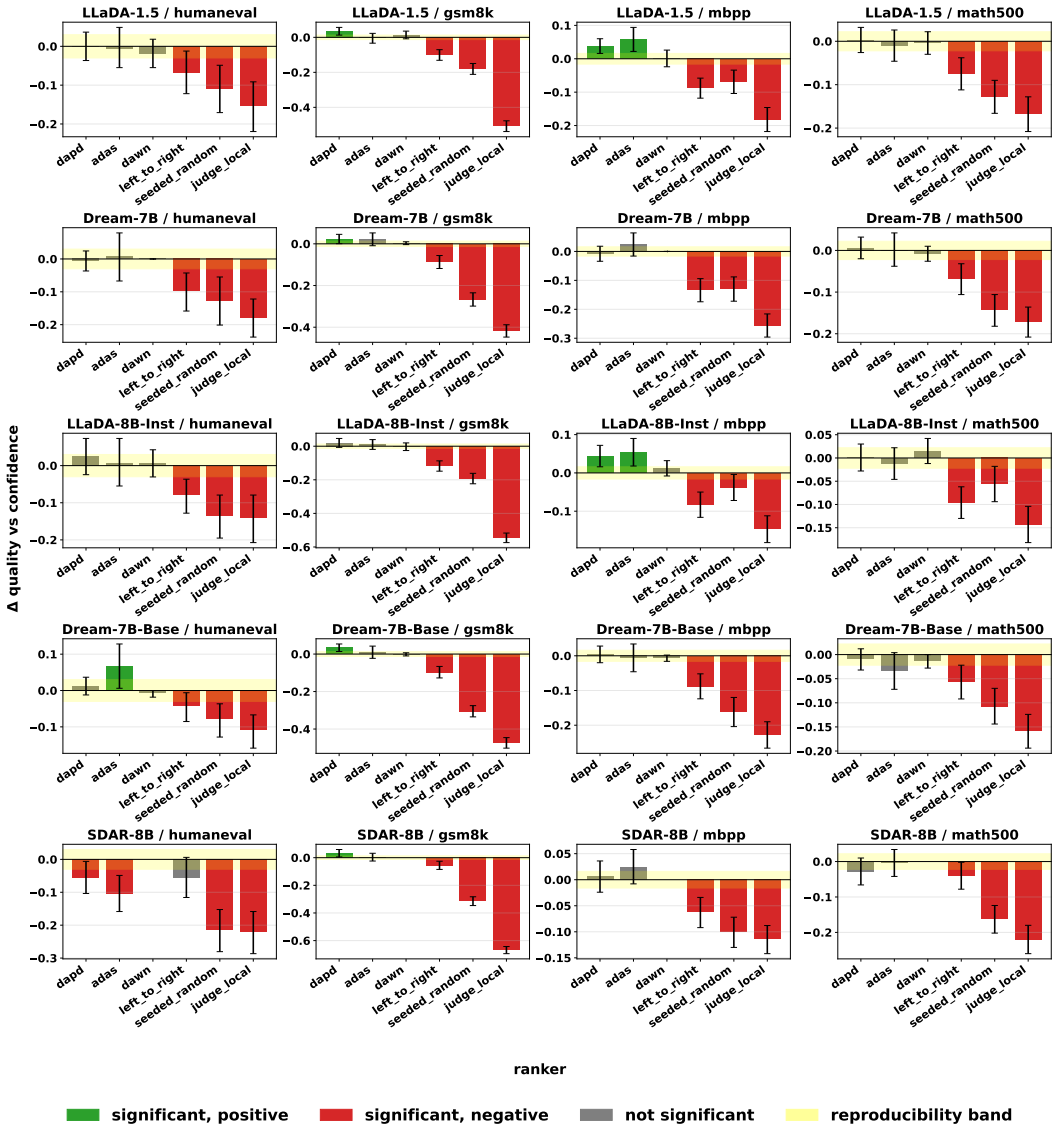

Figure 6: **Order-rule effects relative to confidence at $k = 4$.** Bars show paired accuracy differences with 95% confidence intervals across five models and four tasks. Green and red indicate favorable and unfavorable effects outside the shaded reproducibility band, respectively; gray indicates effects within the band. The learned ranker has the largest negative effect in every panel.

Table 21: **Fixed-pace evidence comparisons at $k = 4$.** Confidence is absolute accuracy (%); other columns are paired percentage-point differences. Attention denotes interaction evidence, temporal denotes distributional stability, and CF denotes length-one counterfactual verification. A star marks an interval excluding zero; a dagger marks an SDAR sensitivity result under policy-specific tie handling.

Task	Conf.	Margin	Entropy	Attention	Temporal	CF
LLaDA-1.5						
HumanEval	22.0	$+6.1^{*}$	-5.5^{*}	-11.0^{*}	-11.0^{*}	$+0.0$
GSM8K	67.8	$+1.5$	-5.4^{*}	-9.8^{*}	-32.2^{*}	$+0.0$
MBPP	20.6	$+3.6^{*}$	-2.6	-6.4^{*}	-4.0^{*}	-0.8
MATH500	25.6	-1.0	-3.6	-8.6^{*}	-13.4^{*}	$+0.0$
Dream-Instruct						
HumanEval	20.1	-1.8	-4.9	-9.1^{*}	-12.8^{*}	$+0.0$
GSM8K	54.2	-2.5	-6.1^{*}	-23.1^{*}	-31.3^{*}	$+0.0$
MBPP	28.8	-2.0	-9.4^{*}	-17.0^{*}	-8.4^{*}	$+0.0$
MATH500	20.6	-2.6	-6.2^{*}	-11.6^{*}	-17.8^{*}	$+0.0$
LLaDA-8B-Instruct						
HumanEval	21.3	$+2.4$	-5.5^{*}	-10.4^{*}	-11.0^{*}	–
GSM8K	63.9	$+1.5$	-7.4^{*}	-14.9^{*}	-39.5^{*}	–
MBPP	18.0	$+6.0^{*}$	-0.8	-6.6^{*}	-4.0^{*}	–
MATH500	19.4	$+1.4$	-6.0^{*}	-5.2^{*}	-13.0^{*}	–
Dream-Base						
HumanEval	11.0	$+2.4$	-3.7	-4.9	-9.1^{*}	–
GSM8K	50.6	-5.5^{*}	-4.9^{*}	-25.5^{*}	-41.8^{*}	–
MBPP	24.8	$+2.2$	-3.8^{*}	-13.6^{*}	-7.2^{*}	–
MATH500	18.0	-2.4	-5.0^{*}	-10.6^{*}	-16.6^{*}	–
SDAR-8B						
HumanEval	25.6	-2.4	$-5.5^{\dagger}$	$-18.3^{*\dagger}$	$-12.8^{*\dagger}$	–
GSM8K	70.7	$+0.5$	$-6.0^{*\dagger}$	$-28.1^{*\dagger}$	$-44.8^{*\dagger}$	–
MBPP	11.6	-0.8	$-1.4^{\dagger}$	$-9.4^{*\dagger}$	$-7.6^{*\dagger}$	–
MATH500	24.8	-0.8	$-1.8^{\dagger}$	$-13.4^{*\dagger}$	$-22.4^{*\dagger}$	–

accumulation over 16 snapshots, and gradient clipping at 5.0. The 8,128 training and 896 validation snapshots are separated by domain and prompt. The resulting datasets contain approximately 106,000 labeled rows per principal backbone.

The ranker achieves pooled validation AUROCs of 0.800 on LLaDA-1.5 and 0.812 on Dream-Instruct. These values measure discrimination over the broadcast group labels across states, whereas deployment requires ranking candidate positions within each state. The training objective does not directly supervise this within-state decision. The experiment therefore provides a counterexample to treating strong pooled discrimination as sufficient evidence of deployment value; it does not imply that learned token-selection policies generally fail.

G.2 Local diagnostics and policy evaluation

We compare three levels of evaluation. First, *reference utility* is the number of consecutive proposed decisions accepted by an exact verifier following a one-token confidence reference. This metric treats every departure from the reference as an error and therefore favors confidence by construction. Second, the one-action diagnostic substitutes one alternative decision into a saved state and completes the trajectory using confidence selection. Third, policy-matched evaluation runs each policy from the initial state to completion, including the state distribution induced by its repeated decisions.

The one-action intervals are approximately ± 4–6 percentage points and leave 19 of 20 comparisons unresolved. These estimates cannot distinguish a small effect that accumulates across decoding from

Table 22: **One-action outcome diagnostics on frozen states.** Values report confidence/arm outcome accuracy (%) and rollout regret (confidence minus arm) in percentage points. Intervals are clustered by prompt over 128 states from 16 prompts per model. Positive regret favors confidence; **bold entries** mark intervals excluding zero.

Arm	Δ utility	Values	Pooled regret	GSM8K-dev	MBPP-dev
LLaDA-1.5					
DAPD	–	49.2 / 49.2	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
ADAS	–	49.2 / 50.0	$-0.8\,[-4.7, +2.3]$	$-3.1\,[-9.4, +0.0]$	$+1.6\,[+0.0, +4.7]$
Margin	–	49.2 / 51.6	$-2.3\,[-4.7, +0.0]$	$-1.6\,[-4.7, +0.0]$	$-3.1\,[-7.8, +0.0]$
JUDGE	–	49.2 / 46.9	$+2.3\,[-3.1, +7.8]$	$+4.7\,[-1.6, +12.5]$	$+0.0\,[-6.2, +7.8]$
DAWN	–	49.2 / 50.0	$-0.8\,[-2.3, +0.0]$	$+0.0\,[+0.0, +0.0]$	$-1.6\,[-4.7, +0.0]$
Random	–	49.2 / 48.4	$+0.8\,[-3.9, +4.7]$	$-1.6\,[-10.9, +6.2]$	$+3.1\,[+0.0, +7.8]$
Dream-Instruct					
DAPD	–	37.5 / 38.3	$-0.8\,[-2.3, +0.0]$	$+0.0\,[+0.0, +0.0]$	$-1.6\,[-4.7, +0.0]$
ADAS	–	37.5 / 37.5	$+0.0\,[-4.7, +4.7]$	$+0.0\,[-4.7, +4.7]$	$+0.0\,[-9.4, +9.4]$
Margin	–	37.5 / 37.5	$+0.0\,[-3.1, +3.9]$	$-1.6\,[-4.7, +0.0]$	$+1.6\,[-4.7, +9.4]$
JUDGE	–	37.5 / 33.6	$+3.9\,[-2.3, +10.2]$	$+3.1\,[-3.1, +7.8]$	$+4.7\,[-6.2, +17.2]$
DAWN	–	37.5 / 37.5	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
Random	–	37.5 / 34.4	$+3.1\,[-2.3, +9.4]$	$+3.1\,[+0.0, +7.8]$	$+3.1\,[-7.8, +15.6]$
LLaDA-8B-Instruct					
DAPD	$+0.04\,[-0.14, +0.32]$	37.5 / 36.7	$+0.8\,[+0.0, +2.3]$	$+1.6\,[+0.0, +4.7]$	$+0.0\,[+0.0, +0.0]$
ADAS	$-0.09\,[-0.31, +0.25]$	37.5 / 38.3	$-0.8\,[-3.9, +2.3]$	$+0.0\,[-4.7, +4.7]$	$-1.6\,[-6.2, +3.1]$
Margin	$\mathbf{-0.72\,[-0.98, -0.47]}$	37.5 / 40.6	$\mathbf{-3.1\,[-6.2, -0.8]}$	$-1.6\,[-4.7, +0.0]$	$-4.7\,[-9.4, +0.0]$
JUDGE	$\mathbf{-1.82\,[-2.02, -1.61]}$	37.5 / 32.8	$+4.7\,[+0.0, +9.4]$	$+1.6\,[-3.1, +6.2]$	$\mathbf{+7.8\,[+1.6, +15.6]}$
DAWN	$-0.22\,[-0.50, +0.15]$	37.5 / 37.5	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
Random	$\mathbf{-1.88\,[-2.12, -1.64]}$	37.5 / 32.0	$+5.5\,[+0.0, +10.9]$	$+4.7\,[-3.1, +12.5]$	$\mathbf{+6.2\,[+1.6, +12.5]}$
Dream-Base					
DAPD	$\mathbf{-0.06\,[-0.12, -0.01]}$	16.4 / 15.6	$+0.8\,[+0.0, +2.3]$	$+0.0\,[+0.0, +0.0]$	$+1.6\,[+0.0, +4.7]$
ADAS	$\mathbf{-0.28\,[-0.41, -0.16]}$	16.4 / 13.3	$\mathbf{+3.1\,[+0.8, +6.2]}$	$\mathbf{+4.7\,[+1.6, +9.4]}$	$+1.6\,[+0.0, +4.7]$
Margin	$\mathbf{-1.05\,[-1.31, -0.80]}$	16.4 / 14.1	$+2.3\,[+0.0, +4.7]$	$+1.6\,[+0.0, +4.7]$	$+3.1\,[+0.0, +7.8]$
JUDGE	$\mathbf{-1.99\,[-2.17, -1.81]}$	16.4 / 14.8	$+1.6\,[-3.9, +7.0]$	$\mathbf{+6.2\,[+1.6, +12.5]}$	$-3.1\,[-10.9, +3.1]$
DAWN	$-0.03\,[-0.07, +0.00]$	16.4 / 16.4	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
Random	$\mathbf{-1.79\,[-1.99, -1.59]}$	16.4 / 14.1	$+2.3\,[-3.1, +7.8]$	$+4.7\,[-1.6, +12.5]$	$+0.0\,[-6.2, +7.8]$
SDAR-8B					
DAPD	$\mathbf{-0.26\,[-0.45, -0.10]}$	0.0 / 0.0	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
ADAS	$\mathbf{-0.09\,[-0.16, -0.03]}$	0.0 / 0.0	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
Margin	$\mathbf{-0.31\,[-0.48, -0.16]}$	0.0 / 0.0	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$
JUDGE	$\mathbf{-2.54\,[-2.90, -2.19]}$	0.0 / 0.8	$-0.8\,[-2.3, +0.0]$	$-1.6\,[-4.7, +0.0]$	$+0.0\,[+0.0, +0.0]$
Random	$\mathbf{-2.72\,[-3.17, -2.25]}$	0.0 / 0.0	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$	$+0.0\,[+0.0, +0.0]$

insufficient precision in the local diagnostic. Reference agreement can also penalize useful deviations: adapted DAPD and margin sometimes reduce accepted depth while improving end-task accuracy. In contrast, the learned ranker performs poorly in all 20 policy-matched comparisons despite its strong pooled AUROC.

Tables 20 and 21 report the complete policy-matched effects, while Table 23 places reference utility beside live accuracy. Table 22 reports the one-action results. Figures 5 and 6 visualize the live effects and their relationship to local reference agreement.

G.3 EXACT REFERENCE VERIFICATION

The serial verifier VJUDGE proposes a nested sequence of confidence-based draft decisions. For each hypothetical prefix, it constructs the corresponding state and independently evaluates the one-token confidence reference. The verifier accepts the longest prefix whose position–token decisions match the reference and otherwise takes a direct reference step. Under identical deterministic computation, accepting only matching decisions preserves the reference trajectory by induction.

Each hypothetical prefix is evaluated through a separate batch-one model pass under the same numerical configuration as the reference. A draft of depth D therefore requires D model evaluations before its accepted prefix is known. This eager computation includes suffix states that may become unnecessary after an earlier rejection, so logical verification calls cannot be interpreted as model-evaluation counts.

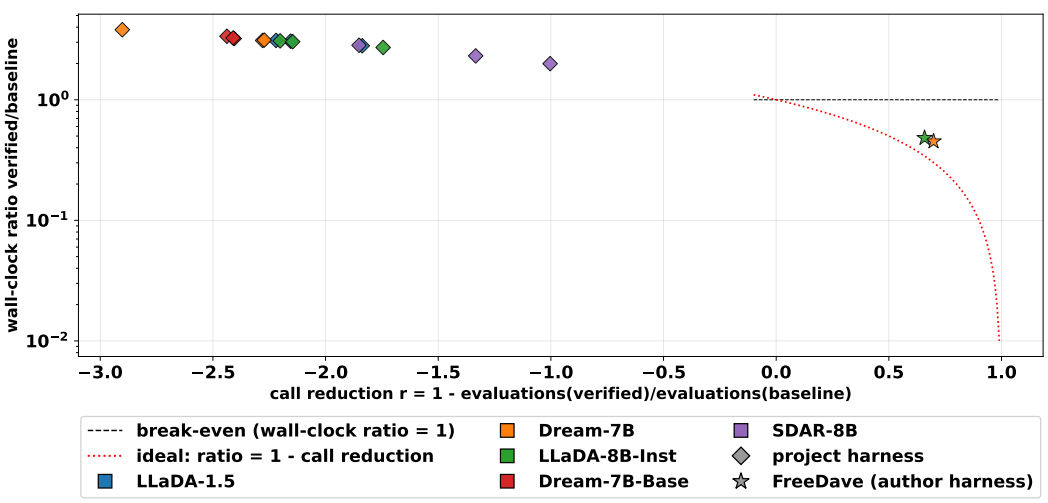

Figure 7: **Runtime versus model-evaluation reduction.** Points compare the serial verifier with its reference in our harness (diamonds) and FREEDAVE with static decoding in the authors' harness (stars). The dashed line marks runtime break-even; the dotted curve shows the ratio expected if runtime scaled only with model-evaluation count.

On full HumanEval, VJUDGE reproduces every reference token array on all five models. Its model-evaluation ratios relative to the reference range from 2.33 to 3.41, with corresponding runtime increases. These measurements characterize the tested serial implementation rather than the computational cost of every possible exact verifier. A restricted-action oracle audit estimates 1.19% additional headroom (95% CI: 0.97–1.40%) within the tested candidate family; it does not bound the potential of other exact-decoding strategies.

G.4 FREEDAVE REPRODUCTION

We evaluate FREEDAVE separately using the authors' implementation (Wu & Zhang, 2025), with its 1,024-token window, block size of 32, caching, and early-exit procedure. On HumanEval-Instruct, its wall-clock ratios relative to static decoding are 0.45 on Dream-Instruct and 0.48 on LLaDA-8B-Instruct. The corresponding accuracy differences are -0.6 percentage points (95% CI $[-3.0, +1.8]$) and $+1.8$ points ($[-0.6, +4.9]$), so neither comparison detects an accuracy change.

The retained outputs are text-identical to static decoding on 128/164 Dream-Instruct items and 96/164 LLaDA-8B-Instruct items. This reproduction therefore establishes faster decoding without a detected accuracy difference, but not token-level identity or exact text agreement on every item. Table 24 reports the fidelity, accuracy, model-evaluation, and runtime results. Figure 7 compares runtime with the corresponding change in model evaluations.

Table 23: **Reference agreement and live policy effects.** Reference utility is the accepted-depth difference from confidence, measured in tokens with state-clustered 95% confidence intervals. Live effects are paired accuracy differences in percentage points. **Bold entries** mark live-accuracy intervals excluding zero.

Model / rule	Task	Δ reference utility	Δ live accuracy
Adapted DAPD			
LLaDA-1.5	HumanEval	$-0.06\,[-0.12, -0.02]$	$+0.0\,[-3.7, +3.7]$
LLaDA-1.5	GSM8K	$-0.09\,[-0.17, -0.03]$	$\mathbf{+3.5\,[+1.3, +5.7]}$
LLaDA-1.5	MBPP	$-0.06\,[-0.12, -0.02]$	$\mathbf{+3.8\,[+1.6, +6.0]}$
LLaDA-1.5	MATH500	$-0.09\,[-0.17, -0.03]$	$+0.2\,[-2.6, +3.2]$
Dream-Instruct	HumanEval	$-0.09\,[-0.17, -0.03]$	$-0.6\,[-3.7, +2.4]$
Dream-Instruct	GSM8K	$-0.05\,[-0.11, +0.00]$	$\mathbf{+2.3\,[+0.1, +4.5]}$
Dream-Instruct	MBPP	$-0.09\,[-0.17, -0.03]$	$-0.8\,[-3.4, +1.8]$
Dream-Instruct	MATH500	$-0.05\,[-0.11, +0.00]$	$+0.6\,[-2.0, +3.2]$
LLaDA-8B-Instruct	HumanEval	$+0.19\,[-0.05, +0.47]$	$+2.4\,[-2.4, +7.3]$
LLaDA-8B-Instruct	GSM8K	$-0.11\,[-0.20, -0.03]$	$+2.0\,[-0.7, +4.6]$
LLaDA-8B-Instruct	MBPP	$+0.19\,[-0.05, +0.47]$	$\mathbf{+4.4\,[+1.6, +7.2]}$
LLaDA-8B-Instruct	MATH500	$-0.11\,[-0.20, -0.03]$	$+0.0\,[-2.8, +3.0]$
Dream-Base	HumanEval	$+0.00\,[-0.08, +0.08]$	$+1.2\,[-1.2, +3.7]$
Dream-Base	GSM8K	$-0.12\,[-0.22, -0.05]$	$\mathbf{+3.3\,[+1.4, +5.4]}$
Dream-Base	MBPP	$+0.00\,[-0.08, +0.08]$	$+0.4\,[-2.0, +2.8]$
Dream-Base	MATH500	$-0.12\,[-0.22, -0.05]$	$-1.0\,[-3.2, +1.2]$
SDAR-8B	HumanEval	$+0.00\,[+0.00, +0.00]$	$\mathbf{-5.5\,[-10.4, -0.6]}$
SDAR-8B	GSM8K	$-0.52\,[-0.75, -0.30]$	$\mathbf{+3.3\,[+0.5, +6.0]}$
SDAR-8B	MBPP	$+0.00\,[+0.00, +0.00]$	$+0.6\,[-2.4, +3.6]$
SDAR-8B	MATH500	$-0.52\,[-0.75, -0.30]$	$-2.8\,[-6.8, +1.0]$
Learned ranker			
LLaDA-1.5	HumanEval	$-1.69\,[-2.02, -1.36]$	$\mathbf{-15.2\,[-22.0, -9.1]}$
LLaDA-1.5	GSM8K	$-1.78\,[-2.05, -1.53]$	$\mathbf{-50.8\,[-53.8, -47.7]}$
LLaDA-1.5	MBPP	$-1.69\,[-2.02, -1.36]$	$\mathbf{-18.2\,[-21.8, -14.6]}$
LLaDA-1.5	MATH500	$-1.78\,[-2.05, -1.53]$	$\mathbf{-16.8\,[-20.8, -12.8]}$
Dream-Instruct	HumanEval	$-1.88\,[-2.14, -1.62]$	$\mathbf{-17.7\,[-23.8, -12.2]}$
Dream-Instruct	GSM8K	$-1.81\,[-2.12, -1.50]$	$\mathbf{-41.8\,[-44.7, -38.8]}$
Dream-Instruct	MBPP	$-1.88\,[-2.14, -1.62]$	$\mathbf{-25.6\,[-29.6, -21.6]}$
Dream-Instruct	MATH500	$-1.81\,[-2.12, -1.50]$	$\mathbf{-17.2\,[-20.8, -13.6]}$
LLaDA-8B-Instruct	HumanEval	$-1.69\,[-1.97, -1.39]$	$\mathbf{-14.0\,[-20.7, -7.9]}$
LLaDA-8B-Instruct	GSM8K	$-1.95\,[-2.20, -1.69]$	$\mathbf{-54.6\,[-57.4, -51.7]}$
LLaDA-8B-Instruct	MBPP	$-1.69\,[-1.97, -1.39]$	$\mathbf{-14.6\,[-18.2, -11.2]}$
LLaDA-8B-Instruct	MATH500	$-1.95\,[-2.20, -1.69]$	$\mathbf{-14.4\,[-18.2, -10.4]}$
Dream-Base	HumanEval	$-1.94\,[-2.19, -1.70]$	$\mathbf{-11.0\,[-15.9, -6.7]}$
Dream-Base	GSM8K	$-2.05\,[-2.34, -1.77]$	$\mathbf{-47.4\,[-50.2, -44.6]}$
Dream-Base	MBPP	$-1.94\,[-2.19, -1.70]$	$\mathbf{-22.8\,[-26.6, -19.0]}$
Dream-Base	MATH500	$-2.05\,[-2.34, -1.77]$	$\mathbf{-15.8\,[-19.4, -12.4]}$
SDAR-8B	HumanEval	$-3.12\,[-3.47, -2.75]$	$\mathbf{-22.0\,[-28.7, -15.9]}$
SDAR-8B	GSM8K	$-1.95\,[-2.27, -1.66]$	$\mathbf{-66.9\,[-69.4, -64.3]}$
SDAR-8B	MBPP	$-3.12\,[-3.47, -2.75]$	$\mathbf{-11.4\,[-14.2, -8.8]}$
SDAR-8B	MATH500	$-1.95\,[-2.27, -1.66]$	$\mathbf{-22.0\,[-26.0, -18.0]}$
Margin evidence			
LLaDA-1.5	HumanEval	$-0.55\,[-0.81, -0.28]$	$\mathbf{+6.1\,[+0.6, +11.6]}$
LLaDA-1.5	GSM8K	$-0.83\,[-1.12, -0.55]$	$+1.5\,[-1.3, +4.3]$
LLaDA-1.5	MBPP	$-0.55\,[-0.81, -0.28]$	$\mathbf{+3.6\,[+0.4, +6.8]}$
LLaDA-1.5	MATH500	$-0.83\,[-1.12, -0.55]$	$-1.0\,[-4.4, +2.6]$
Dream-Instruct	HumanEval	$-0.92\,[-1.23, -0.61]$	$-1.8\,[-7.9, +4.3]$
Dream-Instruct	GSM8K	$-0.62\,[-0.94, -0.34]$	$-2.5\,[-5.6, +0.6]$
Dream-Instruct	MBPP	$-0.92\,[-1.23, -0.61]$	$-2.0\,[-6.4, +2.2]$
Dream-Instruct	MATH500	$-0.62\,[-0.94, -0.34]$	$-2.6\,[-6.2, +1.0]$
LLaDA-8B-Instruct	HumanEval	$-0.48\,[-0.77, -0.20]$	$+2.4\,[-4.3, +9.1]$
LLaDA-8B-Instruct	GSM8K	$-0.95\,[-1.27, -0.64]$	$+1.5\,[-1.4, +4.4]$
LLaDA-8B-Instruct	MBPP	$-0.48\,[-0.77, -0.20]$	$\mathbf{+6.0\,[+2.8, +9.2]}$
LLaDA-8B-Instruct	MATH500	$-0.95\,[-1.27, -0.64]$	$+1.4\,[-2.4, +5.2]$
Dream-Base	HumanEval	$-1.11\,[-1.41, -0.81]$	$+2.4\,[-2.4, +7.3]$
Dream-Base	GSM8K	$-0.98\,[-1.33, -0.66]$	$\mathbf{-5.5\,[-8.6, -2.4]}$
Dream-Base	MBPP	$-1.11\,[-1.41, -0.81]$	$+2.2\,[-1.8, +6.4]$
Dream-Base	MATH500	$-0.98\,[-1.33, -0.66]$	$-2.4\,[-6.0, +1.0]$
SDAR-8B	HumanEval	$-0.09\,[-0.17, -0.03]$	$-2.4\,[-9.2, +4.3]$
SDAR-8B	GSM8K	$-0.53\,[-0.80, -0.30]$	$+0.5\,[-2.3, +3.3]$
SDAR-8B	MBPP	$-0.09\,[-0.17, -0.03]$	$-0.8\,[-3.8, +2.2]$
SDAR-8B	MATH500	$-0.53\,[-0.80, -0.30]$	$-0.8\,[-4.8, +3.2]$

Table 24: **Verification fidelity and computational cost.** Ratios are verifier divided by unverified decoding. Exact denotes matching raw token arrays for VJUDGE and matching decoded text for FREEDAVE. The final column reports the median full-HumanEval runtime ratio over three warmed repetitions.

Model	Task	Δ accuracy [95% CI]	Eval. ratio	Time ratio	Exact	Median time
FreeDave author harness						
Dream-Instruct	HE-Instruct	$-0.6\,[-3.0, +1.8]$	0.30	0.45	128/164 (text)	–
LLaDA-8B-Instruct	HE-Instruct	$+1.8\,[-0.6, +4.9]$	0.34	0.48	96/164 (text)	–
VJUDGE bounded evaluation ($n = 15$)						
LLaDA-1.5	HumanEval	$+0.0\,[+0.0, +0.0]$	3.22	3.11	15/15	–
LLaDA-1.5	GSM8K	$+0.0\,[+0.0, +0.0]$	2.84	2.81	15/15	–
Dream-Instruct	HumanEval	$+0.0\,[+0.0, +0.0]$	3.28	3.10	15/15	–
Dream-Instruct	GSM8K	$+0.0\,[+0.0, +0.0]$	3.90	3.81	15/15	–
LLaDA-8B-Instruct	HumanEval	$+0.0\,[+0.0, +0.0]$	3.20	3.08	15/15	–
LLaDA-8B-Instruct	GSM8K	$+0.0\,[+0.0, +0.0]$	2.74	2.71	15/15	–
Dream-Base	HumanEval	$+0.0\,[+0.0, +0.0]$	3.41	3.22	15/15	–
Dream-Base	GSM8K	$+0.0\,[+0.0, +0.0]$	3.44	3.36	15/15	–
SDAR-8B	HumanEval	$+0.0\,[+0.0, +0.0]$	2.00	1.99	15/15	–
SDAR-8B	GSM8K	$+0.0\,[+0.0, +0.0]$	2.85	2.83	15/15	–
VJUDGE full HumanEval						
LLaDA-1.5	HumanEval	$+0.0\,[+0.0, +0.0]$	3.16	3.06	164/164	3.26
Dream-Instruct	HumanEval	$+0.0\,[+0.0, +0.0]$	3.27	3.13	164/164	3.75
LLaDA-8B-Instruct	HumanEval	$+0.0\,[+0.0, +0.0]$	3.15	3.03	164/164	3.19
Dream-Base	HumanEval	$+0.0\,[+0.0, +0.0]$	3.41	3.26	164/164	3.27
SDAR-8B	HumanEval	$+0.0\,[+0.0, +0.0]$	2.33	2.31	164/164	2.08